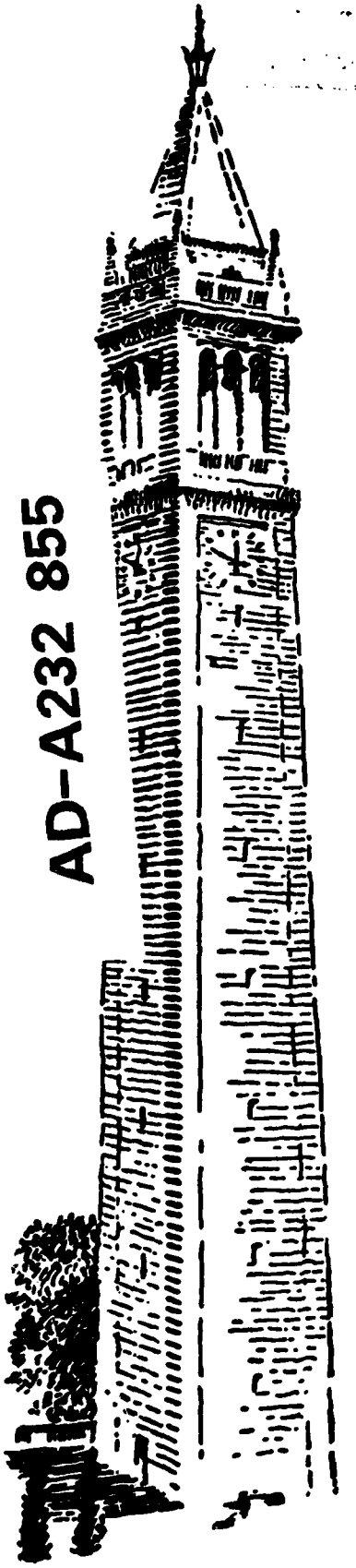


Report No. UCB/ERL 90/1

JOINT SERVICES ELECTRONICS PROGRAM

ANNUAL PROGRESS REPORT
(Contract F49620-90-C-0029)
(6 June 1990 — 28 February 1991)

C. Hu and M. Lieberman

5 March 1991

Prepared for:

Air Force Office of Scientific Research
Bolling Air Force Base
Washington, DC 20332

ELECTRONICS RESEARCH LABORATORY
College of Engineering
University of California, Berkeley, CA 94720

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 5, 1991	Annual 5 Jun 90 - 28 Feb 91

4. TITLE AND SUBTITLE
Annual Report No. UCB/ERL 90/1

5. FUNDING NUMBERS
F49620-90-C-0029
Project/Task: 2305/A9
Program Element: 61102F

6. AUTHOR(S)
Chenming Hu and Michael Lieberman

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Electronics Research Laboratory
253 Cory Hall
University of California at Berekeley
Berkeley, California 94720

8. PERFORMING ORGANIZATION REPORT NUMBER
UCB/ERL 90/1

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research
Building 410
Bolling AFB DC 20332-6448

Program Manager: Dr. Gerald L. Witt, AFOSR/NE

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE,
DISTRIBUTION UNLIMITED

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

This document is the 1990 annual report of research conducted in the Electronics Research Laboratory at University of California at Berkeley under the sponsorship of the Joint Services Electronics Program. The research covers the following project areas: Quantum Electronic Devices: (1-A)Nonlinear Optics in Compound Semiconductors(1-B)Ultrafast Optical Techniques(1-C)Optical Probing of Semiconductor Devices and Interfaces by Electro-Optic and Photo-Elastic Effects; Electronic Devices: (2-A).1 Micron BiCMOS Devices in Bulk and SOI Substrates(2-B)Conductive Oxides and Ferroelectrics for Programmable Devices and(2-C)Insulated-Gate GaAs Field Effect Transistors; Neural Networks and Parallel Computation:(3-A)Stochastic Neural Networks and Application to Signal Processing (3-B) Learning and Generalization by Neural Networks (3-C) Reconfigurable Analog Elements for Neural Nets and (3-D) Architectural Issues in Parallel Computation. There is also a section concerning significant accomplishments.

14. SUBJECT TERMS

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

TABLE OF CONTENTS

Page

PART A -- DIRECTOR'S OVERVIEW 1

PART B -- SIGNIFICANT ACCOMPLISHMENTS

Realization of a Reflector with a Window for Optical Pumping of a Surface Emitting Laser (SELD) 3
Characterizing a Single Hot-Electron-Induced Trap in Submicron MOSFET Using Random Telegraph Noise 5
Ferroelectric Memory Cell with Unlimited Read/Write Cycles 8
Learning Through Logic Synthesis 11
Massively Parallel Analog Geometric Computation Using EEPROMS 12
Fault Tolerance in Feed-Forward Artificial Neural Networks 14

PART C -- INDIVIDUAL WORK UNITS

THEME I -- Quantum Electronic Devices

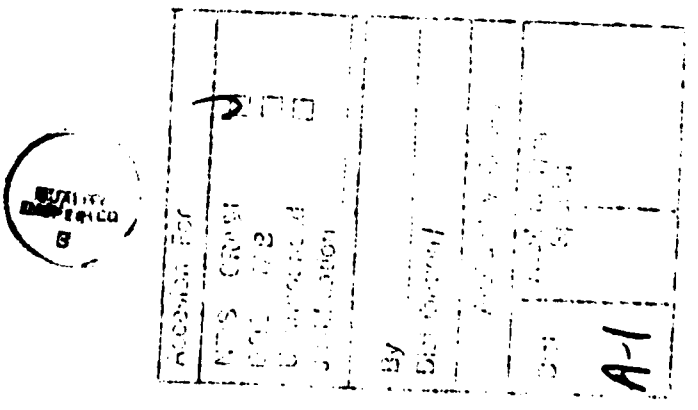
I-A. Nonlinear Optics in Compound Semiconductors 16
I-B. Ultrafast Optical Techniques 21
I-C. Optical Probing of Semiconductor Devices and Interfaces by Electro-Optic and Photo-Elastic Effects 23

THEME II -- Electronic Devices

II-A. 0.1 μm BiCMOS Devices in Bulk and SOI Substrates 28
II-B. Conductive Oxides and Ferroelectrics for Programmable Devices . . . 31
II-C. Insulated-Gate GaAs Field Effect Transistors 35

THEME III -- Neural Networks and Parallel Computation

III-A. Stochastic Neural Networks and Application to Signal Processing 38
III-B. Learning and Generalization by Neural Networks 42
III-C. Reconfigurable Analog Elements for Neural Nets 45
III-D. Architectural Issues in Parallel Computation 47



PART A - DIRECTOR'S OVERVIEW

JSEP continues to play an important role in the electronics research at the University of California, Berkeley. Its emphasis on science and relatively stable funding provide an increasingly rare environment for conducting the more basic research and exploring promising new areas. It also provides a unique opportunity for encouraging collaborative research involving multiple principal investigators or new faculty members. Currently, JSEP at U.C. Berkeley partially supports the research of 10 faculty and 23 graduate students.

Progress is reported here for nine projects under three themes. Under the theme, Quantum Electronic Devices, nonlinear optics in compound semiconductor waveguides are investigated for such applications as signal correlator and spectrum analyzer, ultrafact optical pulse techniques are developed, and electron-optic and photoelastic effects are applied to probe semiconductor devices in interfaces. The theme of Electronic Devices includes three work units. The 0.1 μm Bulk and SOI Devices unit follows and extends the 0.25 μm devices project to explore the new device physic and limitation in future IC devices. Ferroelectrics and conductive oxides are being studied as breakthrough materials for memory devices. Insulated-gate interfaces and dielectrics in GaAs FET are investigated for understanding the limitations and searching for a practical device structure. Theme III integrates various aspects of artificial neural network research into one program. Variable weight devices, especially, EEPROMs, are being developed to implement a novel device concept and ANN learning algorithms. Well developed logic synthesis tools are being used to perform the learning task inherent to the ANN concept. Layered ANN architecture and fault-tolerant ANN are being investigated. The critical issues of parallel computing are studied. Finally, ANN is applied to signal processing.

Several significant accomplishments are highlighted in Part B of this report. A single interface trap generated by hot carriers has been observed and characterized. Transistor current change due to the filling and emptying of a single trap was observed in very small MOSFETs.

A new ferroelectric memory cell overcoming the read-cycle limitation is described. An EEPROM based parallel nearest-neighbor algorithm has the potential for achieving 100X increase in computation time without the need for learning. The fault tolerance of ANN is shown by example to be less than excellent; and a method for improving the fault tolerance is developed.

The enclosed Annual Report Appendix includes copies of 17 published articles, 9 conference papers, 5 papers submitted for publication, and abstracts of 7 Master's/Ph.D. theses.

PART B - SIGNIFICANT ACCOMPLISHMENTS

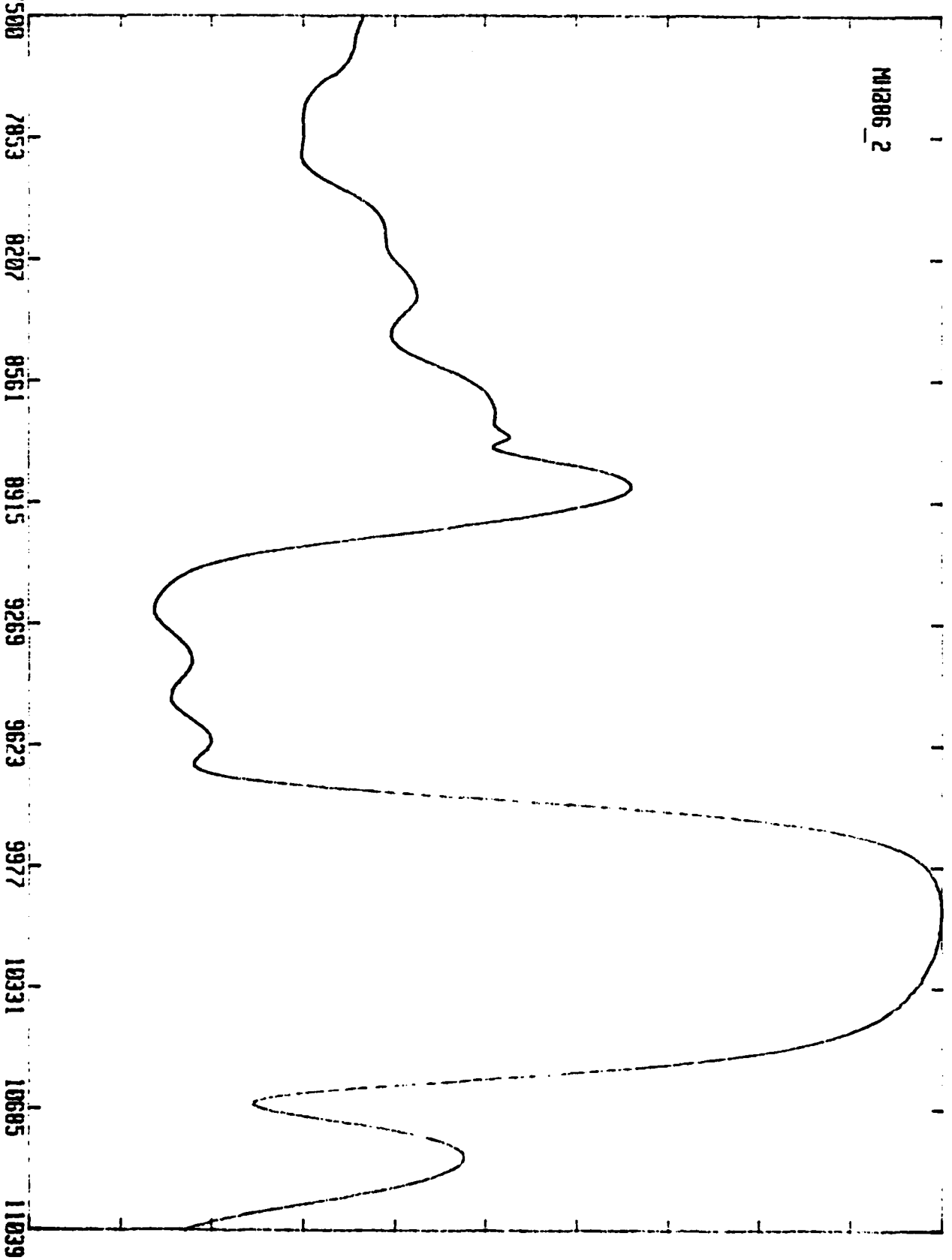
Realization of a Reflector with a Window for Optical Pumping of a Surface Emitting Laser (SELD)

Professor S. Wang with Mark Hadley

A surface emitting laser offers many distinct advantages over an edge emitting laser. These include the possibility of forming a two-dimensional array and the potential of turning on by an optical beam with a speed much faster than achievable by a current pulse. However, an optically pumped SELD requires a specially designed reflector which provides a high reflectivity at the laser wavelength λ_l and a low reflectivity at the pump wavelength λ_p shorter than λ_l. The accompanying figure shows the measured reflectance of such a reflector grown in our laboratory by MBE. The window of low reflectivity from 9,100 to 9,700 *angstroms* is intended for the pump beam to be coupled into the SELD cavity while the region of high reflectivity from 9,900 to 10,300 *angstroms* is intended to provide a high-Q cavity for the laser beam.

The experimentally measured reflectance curve is still far from being ideal, but not because of faulty design. As a matter of fact, the theoretically calculated reflectance curve shows a broad maximum of near-unity reflectivity and a broad window of near-zero reflectivity (or near-perfect transmission). However, because the MBE growth is unexpectedly interrupted, the fabricated structure deviates from the designed structure. With improved computer control of the MBE growth process, we expect to have a reflector of the desired reflectance characteristic. In any case, the experimental result, though it is our first attempt, does confirm the validity of the design concept and does represent the first demonstration of its implementation.

POWER REFLECTIVITY



WAVELENGTH

Characterizing a Single Hot-Electron-Induced Trap in Submicron MOSFET Using Random Telegraph Noise

Professors C. Hu and P.K. Ko with P. Fang

Individual interface traps generated by hot electron stress are observed for the first time [1]. Single trap filling and emptying can cause 0.1% step noise in I_d due to columbic scattering in a very-small-size MOSFET. Trap location (3-10 *angstroms* from interface), time constant, energy and escape frequency are found to be very different from process-induced traps.

The deep-submicron devices used in this study were fabricated using a photoresist-ashing technique [2]. The oxide thickness is 8.6 nm and substrate doping density is $5x10^{17}$ cm^{-3}.

Fig. 1 shows the current noise after hot electron stress at $V_g = 2V$, $V_d = 4.5V$, $I_{sub} = 10$ μA for 10 minutes. It shows the current fluctuation in a deep submicron n-MOSFET with $W_{eff} = 0.5$ μm, $L_{eff} = 0.35$ μm. The striking two-level current fluctuation is due to the filling and emptying of a single interface trap. It is known as the Random Telegraph Noise (RTS) and is observed only when the channel area is small enough to contain only one trap within kT's from the fermi level. This is the first observation of a single hot-electron generated interface trap.

RTS noise can be a useful tool for studying stress-induced interface traps. It is easier to observe stress-induced traps than process-induced traps due to the small stress area and low stress-induced trap density after light stressing. Using RTS as a characterization tool, we found the stress-induced trap to be located closer to the interface, and therefore have a shorter time constant and much stronger influence on scattering and ΔI_d than process-induced traps. Table I lists the time-constants of hot electron-induced interface trap and process-induced, i.e. pre-stress, interface traps. The former are about 50 times shorter than the later.

References

[1] P. Fang, K.K. Hung, P.K. Ko, C. Hu, "Characterizing a Single Hot-Electron-Induced Trap in Submicron MOSFET Using Random Telegraph Noise," *VLSI Technology Symposium Proc.*, pp. 37-36, June 1990.

[2] J. Chung, M.C. Jeng, J.E. Moon, A.T. Wu, T.Y. Chan, P.K. Ko and C. Hu, "Deep Submicron MOS Device Fabrication Using A Photoresist-Ashing Technique," *IEEE Electron Device Lett.*, Vol. EDL-9, p. 186, 1988.

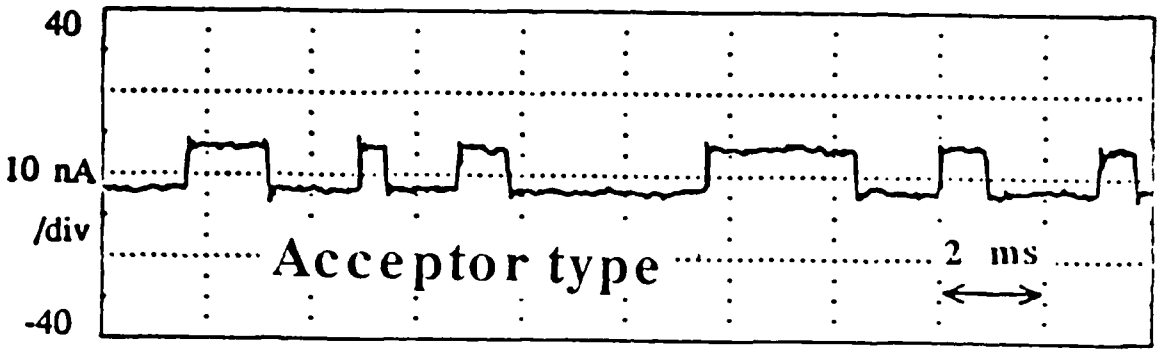


Figure 1. MOSFET (L = 0.35 μm, W = 0.5 μm) current fluctuations due to the filling and emptying of a single interface trap produced by hot-electron stressing.

W / L (μm / μm)	Process-induced		stress-induced		
	τ (ms)	Z (A)	τ (ms)	Z (A)	E_{cox}–E_T(ev)
0.5 / 0.5	257	12	6.1	5.0	3.236
0.5 / 0.35	457	14	7.0	6.4	3.227
0.7 / 0.7	*	*	4.3	3.7	3.296
0.8 / 0.5	330	19	7.3	6.1	3.259
1.0 / 1.0	*	*	8.1	7.0	3.258

Table I. Stress-induced and process-induced trap parameters.

Ferroelectric Memory Cell with Unlimited Read/Write Cycles

Professor C. Hu with R. Moazzami

Single-transistor ferroelectric random access memory (FRAM) has received much attention lately. Ramtron and National Semiconductor have introduced commercial products. For military applications, its radiation hardness and endurance (10^{10} write cycles), superior to those of floating-gate nonvolatile memories, are particularly attractive.

Unfortunately, FRAM makes no distinction between "write" and "read," i.e. the read cycles are also limited to 10^{10} read cycles. 10^{10} read cycles can be consumed in 30 seconds if a cell is read continuously at 33 MHz. DARPA has a large program on FRAM with one major goal being the improvement of the read cycles.

We have invented a new ferroelectric memory cell that overcomes this shortcoming. The cell is read as a DRAM without switching the polarity of the ferroelectric polarization. During "read," the ferroelectric film simply serves as a high-permitivity dielectric in the DRAM capacitor -- enabling a smaller cell size than the state-of-the-art 16 Mb DRAM cell. In the write mode, the memory operates as a FRAM. However, nonvolatile "write" is performed only when the power supply is lost. This scheme is similar to that of NVRAM (nonvolatile RAM). However, a NVRAM cell is twice as large as an SRAM cell and its writing cycle is limited to about 10,000, while the new ferroelectric memory has the size of a DRAM and a write cycle of 10^{10}.

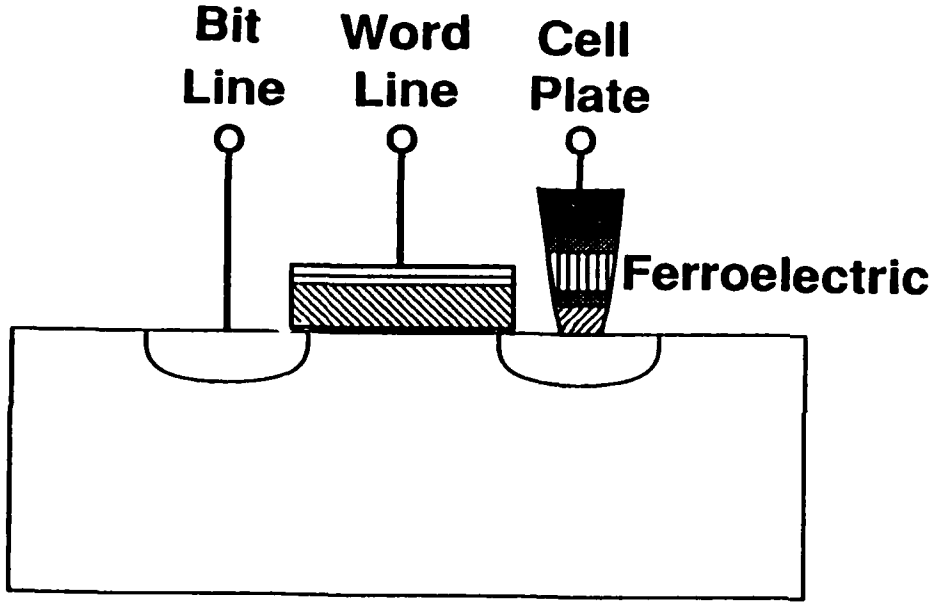
An invention disclosure has been filed. The ferroelectric film characteristics necessary for the cell operation have been studied and reported. Effort is being made to fabricate a working device.

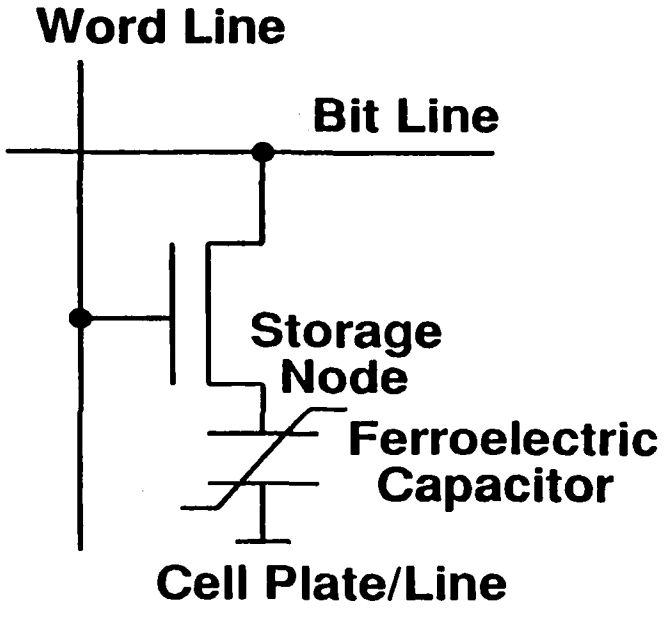
References

[1] R. Moazzami, C. Hu, W.H. Shepherd, "A Ferroelectric DRAM Cell for High Density NVRAM's," *IEEE Electron Device Letters*, pp. 454-456, October 1990.

JSEP HIGHLIGHT

FERROELECTRIC MEMORY CELL WITH UNLIMITED READ/WRITE





- Operates as DRAM but capable of 10^{11} nonvolatile store/recall cycles.
- Eliminates the 10^{11} read/write cycle endurance limit of previous one-transistor ferroelectric memory.
- Entire cell is denser than present DRAMs because ferroelectric film is equivalent to 10Å SiO_2 in the capacitor.

Learning Through Logic Synthesis

Professor Alberto Sangiovanni-Vincentelli with Arlindo Oliveira

We have shown that logic synthesis techniques can be used to derive networks of threshold gates that perform rule induction from examples. Using logic synthesis techniques, we are able to derive networks that perform the required mapping, but are of smaller size than the ones obtained by alternative algorithms. It has been shown that the quality of the induction rules obtained is closely connected to the complexity of the hypothesis generated. Therefore, our ability to generate small networks implies better generalization than the one obtained by alternative techniques.

We have compared the quality of the induction performed by our algorithm with two distint approaches: decision trees and back-propagation. The results obtained in a set of representative examples have shown that the logic synthesis approach outperforms both these methods, both in the number of errors and in the number of examples needed to generate an exact representation of a given concept.

Massively Parallel Analog Geometric Computation Using EEPROMs

A. Kramer, P. K. Ko and Alberto Sangiovanni-Vincentelli

A novel architecture for massively parallel analog computation is presented which uses an EEPROM as the essential computational building block. The proposed architecture is similar in appearance to a memory array, in that the chip stores a set of d-dimensional memories, one per row. In contrast to standard digital memory, each row in the proposed chip is capable of storing in analog either a point, a hypersphere, or a hyperrectangle in d-dimensional space. The other important difference between this architecture and a standard memory chip is that a substantial amount of computation can be performed in parallel directly on all the stored memories. In particular the chip computes geometric relationships based on euclidean distance between the stored memories and a new d-dimensional query point. The relationships the chip is capable of computing include:

- euclidean distance between the query point and all stored points
- euclidean distance between the query point and all stored hyperrectangles
- exclusion of the query point in all stored hyperrectangles
- exclusion of the query point in all stored hyperspheres

Inclusion of some control circuitry and a priority queue allow either the k nearest points/rectangles to the query point or all rectangles/spheres enclosing the query point to be read out of the chip. In addition, the architecture is capable of performing these computations on any subset of the d dimensions, giving it obvious utility as an analog associative memory chip. In fact, the design is similar to a recent digital content addressable memory (CAM) chip [1]. For the CAM task, the proposed analog implementation has several advantages over this digital design including density [$O(10^3)$ times more memory/chip], speed [$O(10^2)$ times as fast] and the fact that it can perform the associative memory function on analog, as well as boolean vectors.

Both the analog storage and the on-chip computation are performed by EEPROMs. Earlier work has show the ability to set the threshold voltage of an EEPROM to an analog value with up to 8-bits of precision [2]. The analog computation is performed by storing points (or rectangles) as the threshold voltages of a row of EEPROMs connected by a common drain and applying the query point as analog voltages on the gates of the same devices. By making use of the inherent I-V characteristics of an EEPROM in saturation ($Id \propto (V_g - V_t)^2$) and taking advantage of current summing, the result of this circuit is a current into each row proportional to euclidean distance squared between the query point and the point stored on that row. The speed [O(μs) settling time] and density [$O(10^6)$ elements per chip, i.e. $10^6/d$ d-dimensional vectors] of the proposed architecture promise to make it a powerful engine for real-world computational tasks such as associative memory and pattern classification.

References

[1] Jon P. Wade and Charles G. Sodini, "A Ternary Content Addressable Search Engine," *IEEE Journal of Solid State Circuits*, Vol. 24, No. 4, pp. 1003-1013, August, 1989.

[2] A. Kramer, et. al, "EEPROM Device as a Reconfigurable Analog Element for Neural Networks," *1989 IEDM Technical Digest*, Beaver Press, Alexandria, Virginia, December 1989.

Fault Tolerance in Feed-Forward Artificial Neural Networks

Reed D. Clay (Professor C. H. Séquin)

The errors resulting from defective units and faulty weights in layered feed-forward ANN's are analyzed, and techniques to make these networks more robust against such failures have been explored. First, using some simple examples of pattern classification tasks and of analog function approximation, we have demonstrated that standard architectures subjected to normal backpropagation training techniques do not lead to any noteworthy fault tolerance. Additional redundant hardware coupled with suitable new training techniques is necessary to achieve that goal.

A simple and general procedure has been found that develops fault tolerance in neural networks: The type of failures that one might expect to occur during operation are introduced at random during the training of the network, and the resulting output errors are used in a standard way for backpropagation and weight adjustment. The result of this training method is a modified internal representation that is not only more robust to the type of failures encountered in training, but which is also more tolerant of faults for which the network has not been explicitly trained.

Ongoing work concerns a more detailed investigation of how the effect of failing hidden units can be mitigated in analog function approximation tasks. We have discovered a promising approach which achieves this goal by tightly controlling the fractional contribution that each hidden unit makes to the linearly summed output value.

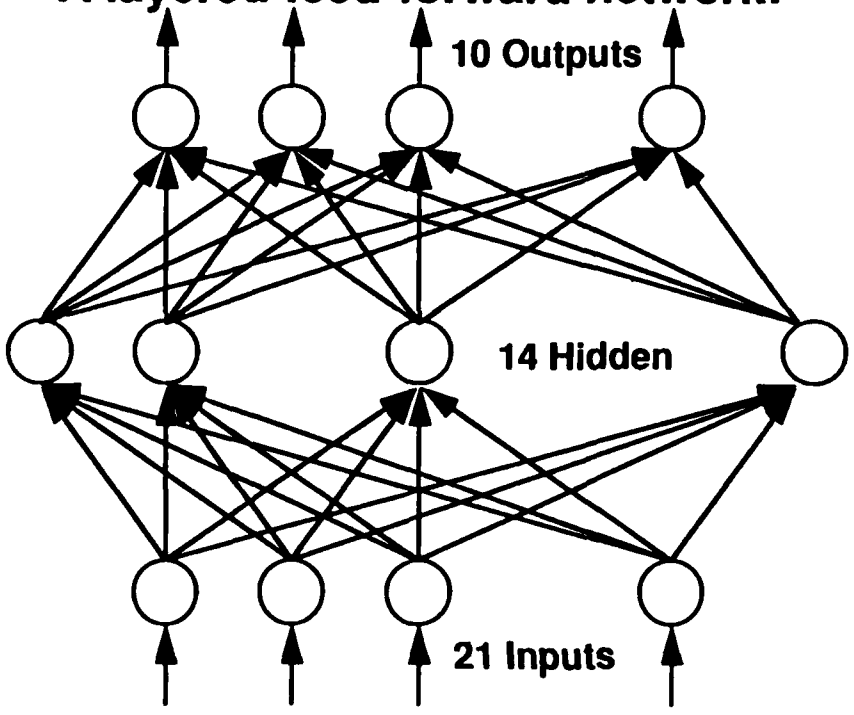
Publications

[1] C. H. Séquin and R. D. Clay, "Fault Tolerance in Artificial Neural Networks", to appear in *Neural Networks, Concepts, Applications and Implementations*," Vol. 4, Antognetti and Milutinovic, eds., Prentice Hall, 1991. Available as *Technical Report* No. 90-031 from International Computer Science Institute, Berkeley.

JOINT SERVICES ELECTRONICS PROGRAM
AT U.C.BERKELEY, EECS DEPARTMENT
PROF. CARLO H. SÉQUIN

FAULT TOLERANCE IN NEURAL NETS

A layered feed-forward network:



RESEARCH FINDINGS :

- Neural nets are **not** a priori fault tolerant. They need to have redundant hardware and they need to be suitably trained.
- Fault tolerance can be obtained by training the network with randomly introduced temporary faults of the type expected during actual operation.
- Prolonged training in this mode leads to stronger fault tolerance, providing:
  - robustness to faults not previously seen and to multiple simultaneous faults.

EXAMPLE :

Net was trained with multiple faults:(#) :	0	1	2	3
Trials required to reach perfection: (#) :	286	676	2066	15118

PART C - INDIVIDUAL WORK UNITS

I-A. Nonlinear Optics in Compound Semiconductors

Professor S. Wang with Patrick Harshman

Our work on nonlinear optics in compound semiconductors during the past year has been focused on two topics: an investigation of (111) strained layer structures and further work on the surface-emitting second-harmonic generation scheme which was developed under JSEP support and reported on last year.

Strained layers grown in the (111) direction have been predicted to possess large built-in electric fields and our goal is to exploit these built-in fields to achieve strongly enhanced second-and third-order nonlinear optical effects. We have made significant progress on the problem of the growth of (111) strained layers, and have demonstrated (111) strained layers which are of high optical quality [1]. More recently, we have studied the low temperature photoluminescence characteristics of these structures and have found evidence which suggests the attainment of self-biased strained quantum wells [2]. This work constitutes the first direct optical evidence of the existence of these built-in electric fields.

Figure 1 shows the 5K photoluminescence spectrum from an AlAs/$Al_{0.5}In_{0.5}As$ multi-quantum well structure grown on a 2° tilted semi-insulating GaAs substrate. We attribute the peak at 8340 *angstroms* (1.487 eV) to a C_1–hh_1 excitonic transition and the peak at 8200 *angstroms* (1.511 eV) to a C_1–hh_2 excitonic transition. In the presence of strain the valence band energy surfaces are described by (1)

$$E_k = Ak^2 + A\varepsilon \pm \sqrt{\xi_k + \xi_{\varepsilon k} + \xi_\varepsilon} \qquad (1)$$

where

$$\xi_k = B^2k^2 + C^2\left(k_x^2k_y^2 + k_y^2k_z^2 + k_z^2k_x^2\right)$$

$$\xi_{\varepsilon k} = Bb\left[3\left(k_x^2\varepsilon_{xx} + k_y^2\varepsilon_{yy} + k_z^2\varepsilon_{zz}\right) - k^2\varepsilon\right] + 2Dd\left(k_xk_y\varepsilon_{xy} + k_yk_z\varepsilon_{yz} + k_zk_x\varepsilon_{zx}\right)$$

$$\xi_\varepsilon = \frac{b^2}{2}\left[\left(\varepsilon_{xx} - \varepsilon_{yy}\right)^2 + \left(\varepsilon_{yy} - \varepsilon_{zz}\right)^2 + \left(\varepsilon_{zz} - \varepsilon_{xx}\right)^2\right] + d^2\left(\varepsilon_{xy}^2 + \varepsilon_{yz}^2 + \varepsilon_{zx}^2\right)$$

Here a, b, and d are the material-dependent deformation potentials and the ε_{ij}'s are the strain tensor elements. For biaxially strained (111) layers $\varepsilon_{xx}=\varepsilon_{yy}=\varepsilon_{zz}$ and $\varepsilon_{xy}=\varepsilon_{yz}=\varepsilon_{zx}$ because of the three-fold symmetry of the (111) plane. Therefore, the change in the energy gap at zone center (k = 0) is

$$\Delta Eg^{\Gamma} = 3a\varepsilon_{xx} \pm \sqrt{3d\varepsilon_{xy}} \tag{2}$$

The strain tensor elements for a strained layer is related to the lattice mismatch through the material elastic tensor elements C_{ij} by (2)

$$E_{xy} = -\left[a_{ref}/a_{layer}-1\right]\left[C_{11}+2C_{12}\right]/\left[4C_{44}+2C_{12}+C_{11}\right] \tag{3}$$

$$\varepsilon_{xx} = -4C_{44}\varepsilon_{xy}/\left[C_{11}+2C_{12}\right]$$

The value of ΔE_g^{Γ} can be calculated using the material parameters compiled by Adachi [3]. The assignment of the PL peak at 8340 *angstroms* to a C_1–hh_1 transition assumes a generally used value of 10 meV for the exciton binding energy. Theoretical investigation is under way to calculate the exciton binding energy in a quantum well, taking into account the anisotropy of effective masses.

Figure 2 shows variation of PL spectra with excitation intensity. Two features of the C_1–hh_1 transition are to be noted. First there is a blue shift at higher excitation intensity I_0. This can be attributed to a screening of the built-in electric field by the optically generated electron-hole pairs. Therefore, the blue shift can be used as evidence of the existence of a built-in electric field. Work is in progress to calculate the wavelength shift caused by the stark effect [4]. Second, the PL spectrum shows saturation at higher I_0. This happens when the carrier concentration becomes sufficiently high to screen the coulomb interaction responsible for the formation of exciton. The C_1–hh_2 transition, on the other hand, does not show any saturation at $I_0 = 145mW$. The reason for non-saturation is being investigated.

We are also presently engaged in the design, growth, and experimental evaluation of an improved performance surface-emitting second-harmonic generator. Our approach is to use

asymmetric quantum wells with near-resonant intersubband transition energies to achieve a second-order susceptibility which is much larger than that of the previously used bulk GaAs. The quantum well scheme has the additional advantage of offering the potential for phase-matching in the propagation direction of the surface-emitted second-harmonic signal.

References

[1] G.L. Bir and G.E. Pikus, *Symmetry and Strain-induced Effects in Semiconductors*, Wiley, 1974.

[2] J.F. Nye, *Physical Properties of Crystals*, Oxford University Press, 1964.

[3] S. Adachi, *J. Appl. Phys.*, Vol. 53, p. 8777, 1982 and Vol. 58, RI, 1985.

[4] P.J. Harshman, "Studies of Growth and Photoluminescence Characteristics of (111) Strained Layers," Master's Project, University of California, Berkeley, California, December 1990.

Publications

[1] P.J. Harshman, K.J. Malloy, J. Walker, J.S. Smith, and S. Wang, "MBE Growth of High Quality (111)B GaAs, GaInAs, and AlInAs," *Mat. Res. Soc. Symp. Proc.*, Vol. 198, pp. 265, 1990.

[2] P.J. Harshman, "Investigation of Growth and Photoluminescence Characteristics of (111) Strained Layers," Master's Project, University of California, Berkeley, California, December, 1990.

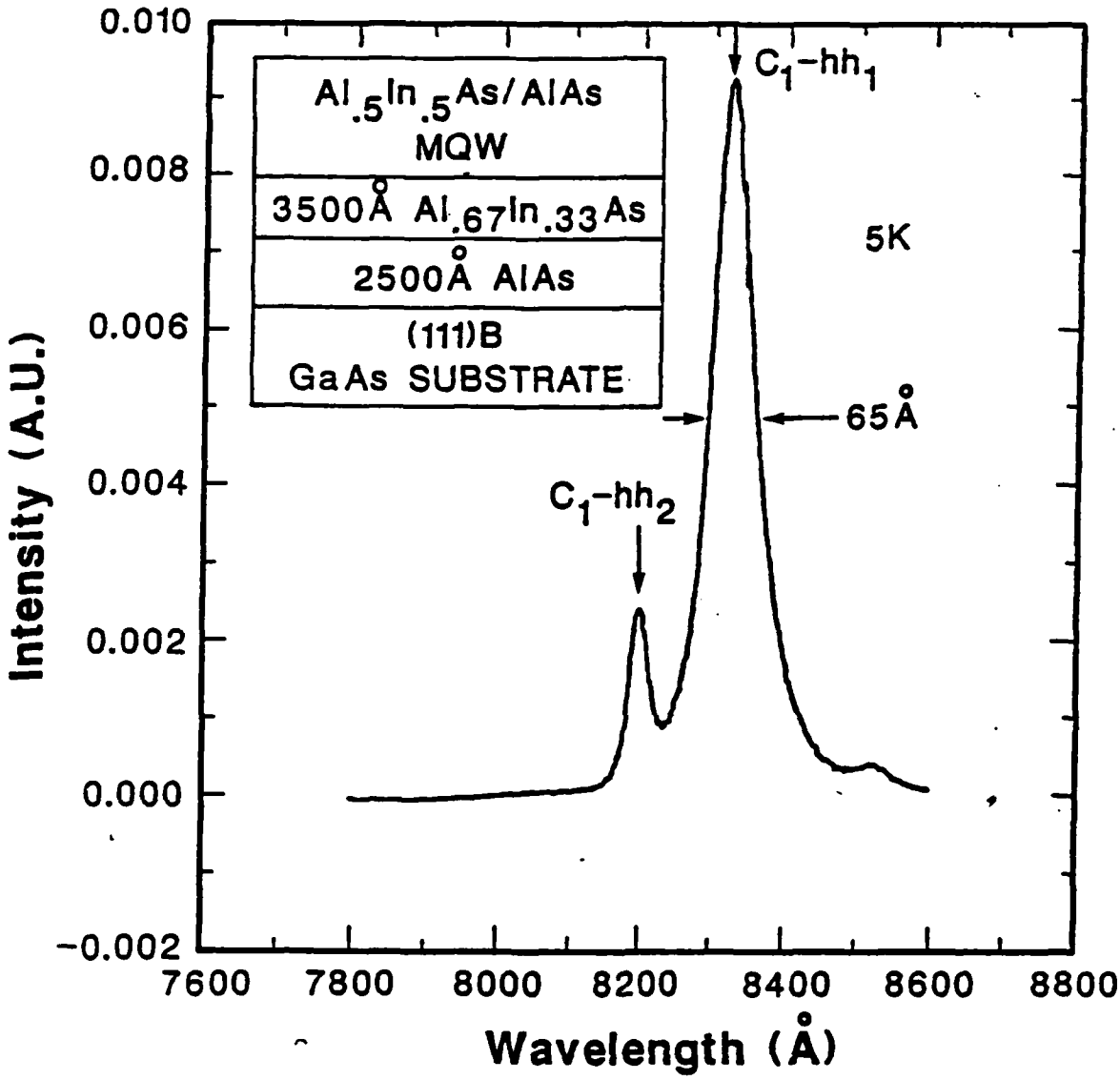
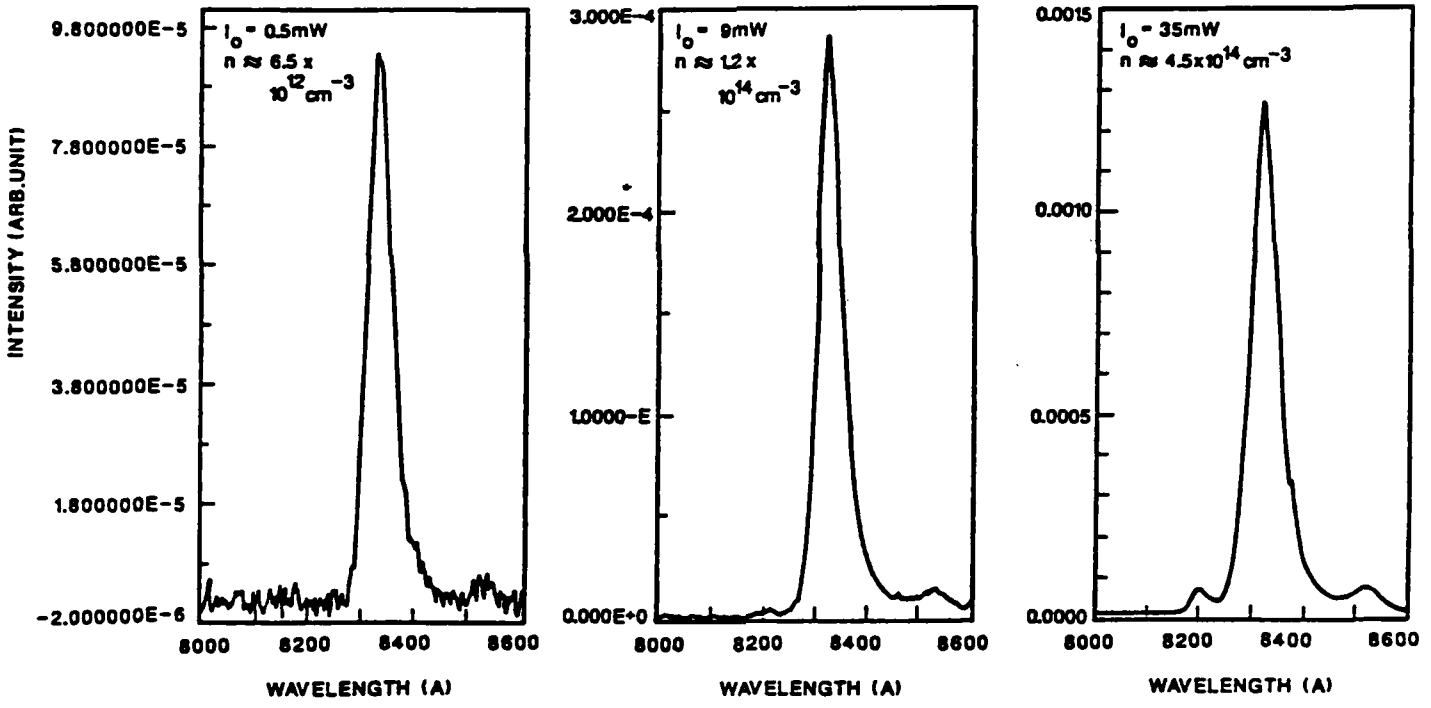

0.010
0.008
0.006
0.004
0.002
0.000
-0.002
Intensity (A.U.)
7600
7800
8000
8200
8400
8600
8800
Wavelength (Å)
Al.5In.5As/AlAs MQW
3500Å Al.67In.33As
2500Å AlAs
(111)B GaAs SUBSTRATE
C1-hh1
C1-hh2
65Å
5K

FIGURE 1


INTENSITY (ARB.UNIT)
9.800000E-5
7.800000E-5
5.800000E-5
3.800000E-5
1.800000E-5
-2.000000E-6
I_o = 0.5mW
n ≈ 6.5 x 10^{12} cm^{-3}
3.000E-4
2.000E-4
1.0000-E
0.000E+0
I_o = 9mW
n ≈ 1.2 x 10^{14} cm^{-3}
0.0015
0.0010
0.0005
0.0000
I_o = 35mW
n ≈ 4.5x10^{14} cm^{-3}
8000
8200
8400
8600
WAVELENGTH (A)

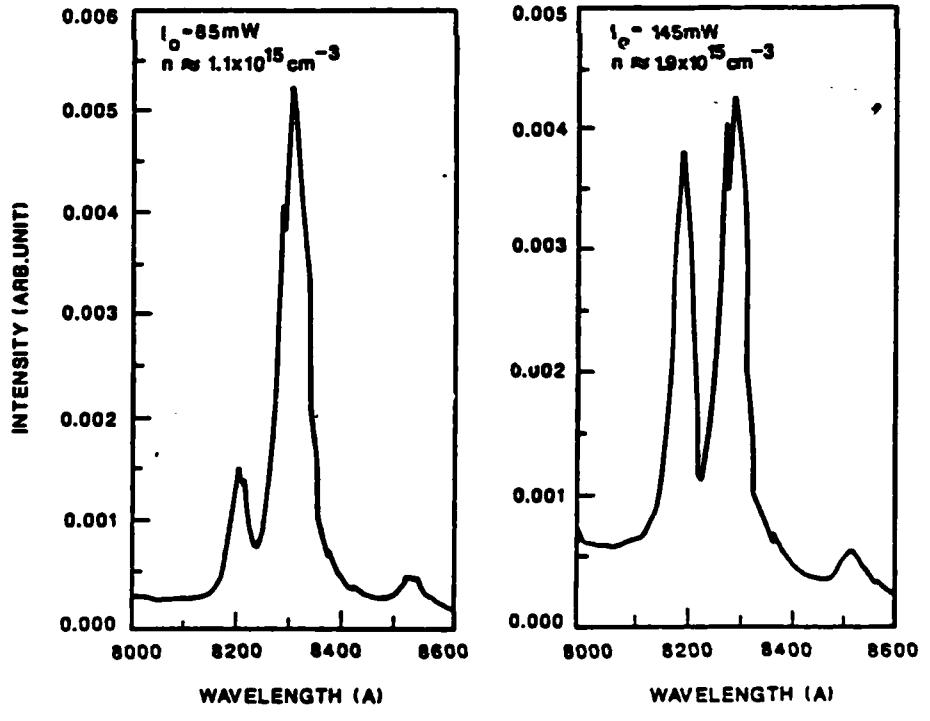

INTENSITY (ARB.UNIT)
0.006
0.005
0.004
0.003
0.002
0.001
0.000
I_o = 85mW
n ≈ 1.1x10^{15} cm^{-3}
0.005
0.004
0.003
0.002
0.001
0.000
I_o = 145mW
n ≈ 1.9x10^{15} cm^{-3}
8000
8200
8400
8600
WAVELENGTH (A)

FIGURE 2

I-B. Ultrafast Optical Techniques

Professor John Stephen Smith with Hong Lin, Jeff Walker, Sol Dijaili, Gordon Wilson and James Yeh

We have now demonstrated low threshold, moderately high power surface emitting lasers using our Phase locked Molecular Beam Epitaxy technique. We are currently emphasizing work on higher power versions of this laser structure. In addition we plan to use multiquantum well saturable absorbers embedded in an MBE grown Bragg layer distributive reflector to passively mode lock the Ti:Sapphire laser. A version of this technique was first demonstrated by Keller, *et al.* A wavelength tunable short pulse system will allow us to characterize the saturable absorber/mirror.

We have recently demonstrated the usefulness of cross-phase modulation in a semiconductor amplifier to modulate the properties of short pulses. In particular, we have demonstrated the removal of an adiabatic chirp from ultra short pulses using cross-phase modulation. A simple expression for the chirp imparted on a weak signal pulse by the action of a strong pump pulse has been derived. A novel dispersive technique for characterizing the resulting nonlinear chirp was introduced and used in the experiment. A maximum frequency excursion of 16 GHz due to the cross-phase modulation was measured. A value of 6 was found for α_{xpm} which is a factor for characterizing the cross-phase modulation in a similar manner to the conventional linewidth enhancement factor, α.

We are investigating the use of frequency up-conversion to extract and demultiplex high bandwidth and time multiplexed signals from optical fibers. The technique uses a synchronized local optical pulse source which is optically mixed with the signal, and the sum frequency is detected with a PIN device or other detector. This would allow one channel of N to be decoded where N is the ratio of the speed of the local optical pulse to the speed of the detector at the sum frequency. The sum frequency conversion efficiency can be high, with sufficient

local pulse power. We have constructed a demonstration two channel system. The time resolution of this technique and the noise performance has been measured. An interesting extension of this work will be the use of a partially phase matched technique in which the nonlinear interaction takes place in a waveguide, and the sum frequency is generated as a wave propagating out from the waveguide at an angle. The dispersion of the waveguide will then sweep the local pulse temporally through the signal, converting the time domain into spatially separated, low frequency signals. Thus, with a single local pulse, the waveguide device, and an array of simple detectors, many channels can be decoded at once.

Publications/Papers

[1] S.P. Dijaili, A. Dienes and J.S. Smith, "ABCD-Matrices for Dispersive Pulse Propagation," *IEEE Journal of Quantum Electronics*, Vol. 26, No. 6, June 1990.

[2] J. Walker, K. Malloy, and S. Wang and J.S. Smith, "Precision Bragg Reflectors Fabricated by Phase-locked Epitaxy," *Applied Physics Letters*, Vol. 56, No. 25, pp. 2493-2495, June 1990.

[3] H. Lin, S. Dijaili, and J.S. Smith, "Ultra High Speed Time Division Demultiplexing by Optical Parametric Interaction" presented at the Conference on Lasers and Electro-Optics, Anaheim, California, May 1990.

[4] S.P. Dijaili, J.M. Wiesenfeld, G. Raybon, C.A. Burrus, A. Dienes, J.S. Smith, J.R. Whinnery, "Cross Phase Modulation in a Semiconductor Laser Amplifier Determined by a Dispersive Technique," submitted to *Applied Physics Letters*.

I-C. Optical Probing of Semiconductor Devices and Interfaces by Electro-Optic and Photo-Elastic Effects

Professor S. Wang with Mark Hadley

One important development in GaAs technology is the discovery of a low-temperature GaAs buffer layer [1] which has eliminated backgating effect. Even though the LTBL has a very low level of response to excitation by visible light due to extremely short carrier lifetime, it can respond to below-gap photo-excitation by lifting electrons from deep-level traps. As a matter of fact, such a study can lead to elucidation of the physical mechanism responsible for the semi-insulating property. We plan to use tunable Ti-sapphire laser as the excitation source to study the dynamics of carrier de-trapping and trapping. As a preparatory step, we have started to grow LT buffer layers under different growth conditions. Figure 1 shows the quality of the grown layers as functions of layer thickness and substrate temperature at a constant As/Ga flux ratio. The solid line indicates demarkation between films of good and bad morphology. We plan to take and examine TEM micrographs to determine the sizes and density of As precipitates [2,3].

While we are waiting for LT GaAs films, we have come up with a novel and potentially important idea of making an optical pumping surface emitting laser. Periodic-layered structures commonly used for surface emitting laser diode (SELD) exhibit a reflectance curve consisting of a main stop band and symmetric diffraction side-lobes. This type of reflectance characteristic limits the percentage of optical power which can be coupled into the laser cavity. For optical pumped SELD, we need a window in the reflectance curve through which a large percentage of pumping power can be coupled into the active region of the laser cavity.

We have looked at ways to increase the efficiency of an optically pumped SELD. Theoretically, the maximum amount of power absorbed from a pump beam is given by

$$\frac{P_{absorbed}}{P_{incident}} = 1 - \frac{e^{-\alpha l}(1 - R)}{[1 - Re^{-2\alpha l}]} \qquad (1)$$

where $e^{-\alpha l}$ is the loss of the pump beam in the gain region and R is the reflectivity of the lower mirror. Some typical numbers are $e^{-\alpha l} = 0.9$ and $R = 0.98$. Using these numbers in Eq. (1) shows that 91% of the pump beam can be absorbed. This is much higher than reported in any publication. In order to achieve this maximum absorption it is necessary to modify the upper mirror structure to match the power into the gain region.

For the specific case of large loss, the problem reduces to making the top mirror transparent at the pump wavelength. A computer program was designed to make a "pump window" in the mirror without reducing the peak reflectivity. A theoretical plot of such a mirror is shown in Figure 2. Such a mirror was grown using MBE with good results. The experimental reflectance curve for the MBE-grown is shown in Figure 3. Note the appearance of a broad window of low reflectance in the 9,100 to 9,700 *angstroms* region. Due to an unexpected interruption and change in condition in MBE growth, the experimental curve deviates somewhat from the theoretical curve. However, the result does confirm the validity of the concept and represents the first demonstration of the concept.

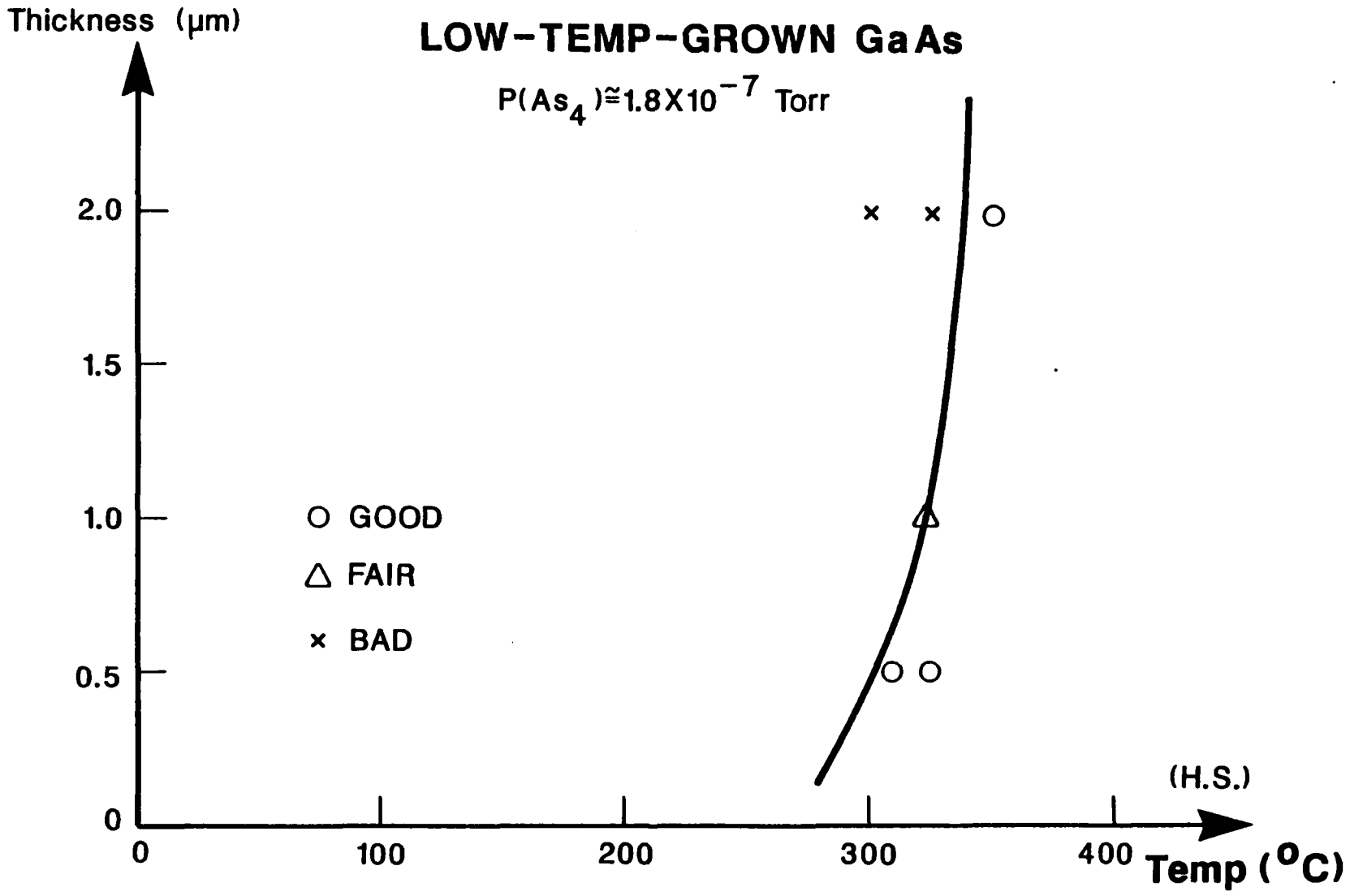

LOW-TEMP-GROWN GaAs
P(As4)≅1.8X10−7 Torr
Thickness (μm)
2.0
1.5
1.0
0.5
0
○ GOOD
△ FAIR
× BAD
0
100
200
300
400
(H.S.)
Temp (°C)

FIGURE 1

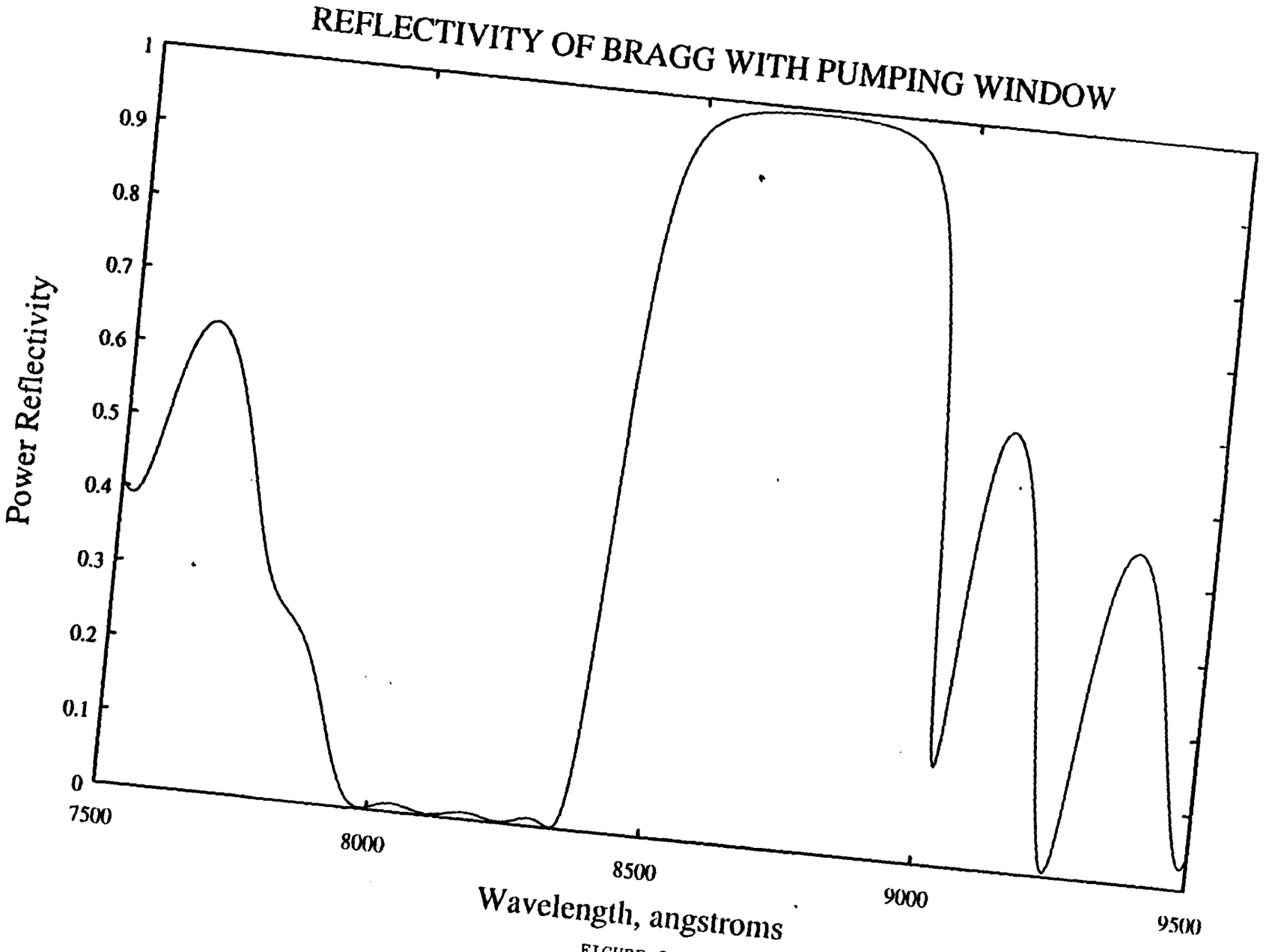

REFLECTIVITY OF BRAGG WITH PUMPING WINDOW
Power Reflectivity
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
7500
8000
8500
9000
9500
Wavelength, angstroms

FIGURE 2

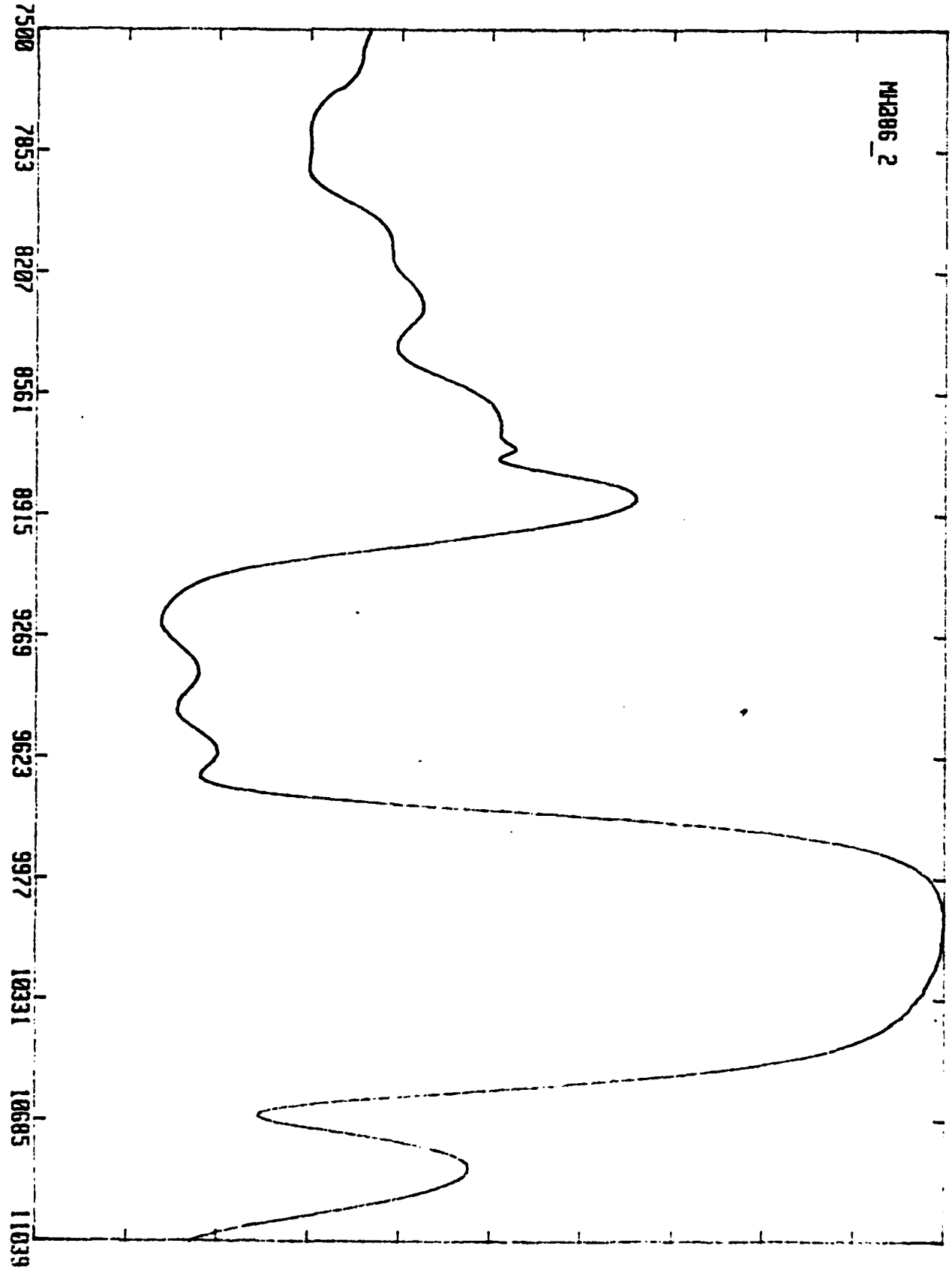

MW286_2
7500
7853
8207
8561
8915
9269
9623
9977
10331
10685
11039

FIGURE 3

II-A. 0.1 μm BiCMOS Devices in Bulk and SOI Substrates

Professors C. Hu and P.K. Ko with J. Chung, P. Nee and F. Assaderaghi

During this period we have designed and generated two new mask sets for 0.1 μm CMOS and lateral bipolar transistors. One set was designed with SOI (silicon on insulator) devices in mind. Development of the lithography and etching techniques necessary for fabricating 0.1 μm devices has begun. In the meantime we have characterized submicron MOSFETs down to 0.2 μm channel length, and begun to investigate the SOI devices. Eight papers resulting from this work unit have been published in this period.

The two test masks were designed to test the feasibility of producing 0.1 μm channel length or lateral base width using a novel lithography technique previously developed with JSEP support. An I-line stepper is used to produce 0.6 μm photoresist lines. The developed photoresist pattern is isotropically etched in oxygen plasma in a manner often called photoresist ashing until the desired linewidth is achieved. This technique has been used successfully by us to fabricate 0.25 μm devices under JSEP sponsorship. Those devices set the world record for room-temperature silicon device speed at 22ps. We have just completed one test run aimed at producing 0.1 μm devices. The electrical results are unclear because of an ohmic contact problem. The samples are currently being studied using high resolution SEM.

One of the mask sets has SOI devices in mind. Both N-channel and P-channel MOSFETs are included. A unique feature is that several novel structures, having the MOSFET body, either shorted to the source or separately contacted. The SOI MOSFET body is usually floating. The floating body is suspected of causing low breakdown voltage and of preventing measurement of the substrate current -- a powerful tool for studying the high field effects in deep submicron MOSFETs. These structures will eliminate these problems. Many lateral bipolar transistor structures are also designed. We believe that complimentary (NPN and PNP) lateral bipolar devices will be very attractive on SOI substrates.

We have found a noise peak in SOI MOSFETs and used the noise to study the SOI Si/SiO_2 interfaces [1]. One important conclusion is that the bottom interface in SIMOX device has no higher interface trap density than state-of-art gate oxides.

For the first time a single hot-carrier generated interface trap has been observed and characterized [2]. In a sufficiently small MOSFET, e.g., 0.2 μm x 0.5 μm, a single interface trap can cause clear step-function changes in the source-drain current as the trap is filled and emptied. The height and frequency of the step function as well as its gate voltage dependence allowed us to thoroughly characterize a single trap.

The effect of hot-electron degradation of submicron devices on analog MOSFET performance was studied [3]. Although the effect of device degradation on digital circuit has been widely studied, ours is the first study of its impact on analog circuit reliability. We concluded that many analog circuits are much more susceptible to hot electron induced degradation than the typical digital circuits. N and P MOSFET degradation by hot carriers was also studied [4,5] and a gate current model was presented for PMOSFETs for the first time. We have also concluded a study of the effects of source/drain series resistance on deep submicron devices [6]. Reduction in the measured saturation drain current relative to the ideal saturation current (R_{sd} = 0.0 Ω μm) is about 4% for L_{eff} = 0.7 μm and T_{ox} = 15.6 nm, and 10% for L_{eff} = 0.3 μm and T_{ox} = 8.6 nm. Reduction of current in the linear regime and reduction of the simulated ring oscillator speed are both about 3 times higher. Silicidization of the source/drain is estimated to eliminate as much as 50% of the performance degradation.

Finally, two studies of submicron GaAs MESFETs initiated last year have been completed. It was found that backgating in GaAs MESFETs at high drain voltages can be significantly reduced by properly adjusting the EL-2 center concentration. The reduction is due to the compensation of the negative space charge at the channel substrate interface by holes generated by impact ionization in the MESFET channel [7]. A simple model is presented for

the negative drain current transients observed in GaAs MESFETs when subjected to ionizing radiation [8]. The two dominant mechanisms are proposed to be electron trapping under the Schottky gate and in the neutral semi-insulating substrate. The model is suitable for the design and evaluation of radiation-resistant GaAs MESFET integrated circuits using common electrical simulators such as SPICE3.

Publications

[1] J. Chen, P. Fang, P.K. Ko, and C. Hu, "Noise Overshoot at Drain Current Kink in SOI MOSFET," *1990 IEEE SOS/SOI Technology Conference Proceedings*, p. 40-41, October 2-4, 1990.

[2] P. Fang, K.K. Hung, P.K. Ko, and C. Hu, "Characterizing a Single Hot-Electron-Induced Trap in Submicron MOSFET Using Random Telegraph Noise," *Digest of Technical Papers of Symposium on VLSI Technology*, Honolulu, Hawaii, pp. 37-38, June 1990.

[3] J.E. Chung, K.N. Quader, C.G. Sodini, P.K. Ko, and C. Hu, "The Effects of Hot-Electron Degradation on Analog MOSFET Performance," *Technical Digest of IEEE International Electron Devices Meeting*, pp. 553-556, December 1990.

[4] J. Chung, M-C. Jeng, J. Moon, P.K. Ko, and C. Hu, "Low Voltage Hot Electron Degradation in Deep Submicron MOSFETs," *IEEE Trans. on Electron Devices*, Vol. 37, No. 7, pp. 1651-1657, July 1990.

[5] T.C. Ong, P.K. Ko, and C. Hu, "Hot-Carrier Current Modeling and Device Degradation in Surface Channel P-MOSFET," *IEEE Trans. on Electron Devices*, Vol. 37, No. 7, pp. 1658-1666, July 1990.

[6] M.C. Jeng, J.E. Chung, P.K. Ko, and C. Hu, "The Effects of Source/Drain Resistance on Deep Submicron Device Performance," *IEEE Trans. on Electron Devices*, Vol. 37, No. 11, pp. 2408-2410, November 1990.

[7] P. George, K. Hui, P.K. Ko, and C. Hu, "The Reduction of Backgating in GaAs MESFET's by Impact Ionization," *IEEE Electron Device Letts.*, pp. 434-436, October 1990.

[8] P. George, P.K. Ko, C. Hu, "Model for Photo-Induced Long-Term Drain Current Transients in GaAs MESFETS," *Int. J. of Electronics*, Vol. 68, No. 5, pp. 721-728, October 1990.

II-B. Conductive Oxides and Ferroelectrics for Programmable Devices

Professor C. Hu with Reza Moazzami and H. Shin

The large charge storage density requirement for future generations of DRAMs has generated significant interest in high dielectric constant materials such as tantalum pentoxide and yttrium oxide. However, because of the lower dielectric breakdown strengths of these materials, the net gain in charge storage density has been a factor of two or three at best. Recently, nonvolatile memory cells exploiting the large polarization and ferroelectric hysteresis loops of materials such as lead zirconate titanate ($PbZr_xTi_{1-x}O_3$, commonly called PZT) have been proposed. However, because these memories suffer from fatigue, a gradual loss of polarization following repeated read/write cycling, ferroelectric materials have also been considered as a direct replacement for oxide/nitride/oxide structures in conventional volatile DRAMS. In this case, the ferroelectric is not cycled between the two polarization states during read/write operation thus possible avoiding significant fatigue. We have proposed a ferroelectric nonvolatile RAM (FNVRAM) which normally operates as a conventional DRAM yet also exploits the hysteresis loop of ferroelectric materials for nonvolatile operation [1,2]. The relevant properties of the PZT films were studied [1,2,3].

The FNVRAM cell is a simple one-transistor DRAM cell with a ferroelectric capacitor as shown in Fig. 1. A conductive diffusion barrier is required as the storage node contact to prevent interdiffusion of silicon with the ferroelectric material during high-temperature annealing. Two different bias schemes for the operation of the FNVRAM cell were described. In one scheme, the cell plate is always held at half of the supply voltage ($V_{DD}/2$). During DRAM operation, the storage node is at either $V_{DD}/2$ or V_{DD} such that the ferroelectric capacitor is not cycled between opposite polarization states. Upon command or power failure, a nonvolatile store operation is executed: the state of the cell is read and written back as one of the two permanent polarization states of the ferroelectric film. If the DRAM datum is *zero*,

i.e., the storage node is at $V_{DD}/2$, the word line is selected and the bit line is grounded. The ferroelectric is now polarized in one direction (nonvolatile *zero).* The recall operation is performed similarly to a DRAM read: the remanent polarization of the ferroelectric film is sensed and restored as one of the two DRAM states.

In this manner, the ferroelectric film is only cycled between opposite polarization states during nonvolatile store/recall operations, not during DRAM read/write operations. Even after 10^{10} store/recall cycles (corresponding to 10 store/recall cycles per second for 30 years), there is sufficient detectable ferroelectric polarization. Therefore, fatigue from store/recall cycling is not a serious limitation to the nonvolatile operation of this cell. Since the ferroelectric polarization is not reversed during DRAM read/write operation, there is almost no loss in nonvolatile polarization even after 10^{10} read/write cycles. At this rate, the FNVRAM cell is expected to tolerate orders of magnitude higher nonswitching read/write cycles than the 10^{12} switching cycles demonstrated for PZT films. A loss in the detectable polarization is observed even during DRAM operation. However, a lower limit for the available polarization can be obtained from the small-signal capacitance of the ferroelectric film. For the unoptimized 4000-*angstrom* PZT films studied here, this lower limit is 60 $fC/\mu m^2$ for a 3-V power supply equivalent to a 17-*angstrom* silicon dioxide film. The resistivity and endurance properties of ferroelectric films can be optimized by modifying the composition of the film. This cell can be the basis of a very high-density NVRAM with practically no read/write cycle limit and at least 10^{10} nonvolatile store/recall cycles.

We have completed a study of the enhanced conductivity of oxides grown on heavily doped substrates. It was found that the enhanced conductivity can be attributed to the thinning of oxide at the field edge. This is a rather surprising finding which makes it doubtful that such oxides will be useful for nonvolatile memory applications. On the other hand, it has shown a way of avoiding such enhanced conduction in such cases as oxides grown on heavily doped

substrates for switched-capacitor circuits and radiation-hard field isolation. A publication is being prepared.

References

[1] R. Moazzami, C. Hu, and W.H. Shepherd, "A Ferroelectric DRAM Cell for High Density NVRAM's," *IEEE Electron Device Letts.*, pp. 454-456, October 1990.

[2] R. Moazzami, C. Hu, and W.H. Shepherd, "Endurance Properties of Ferroelectric PZT Thin Films," *Technical Digest of IEEE International Electron Devices Meeting*, pp. 417-420, December 1990.

[3] R. Moazzami, C. Hu, and W.H. Shepherd, "A Ferroelectric DRAM Cell for High Density NVRAMs," *Digest of Technical Papers of Symposium on VLSI Technology*, Honolulu, Hawaii, pp. 15-16, June 1990.

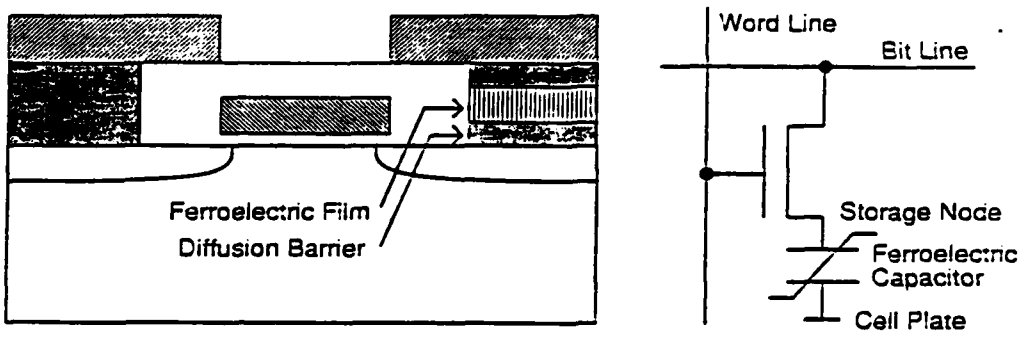


Fig. 1. The FNVRAM cell is a conventional DRAM cell with a ferroelectric capacitor dielectric. Since the ferroelectric material has a very large polarization, it is possible to incorporate the capacitor in the contact hole of the select transistor.

II-C. Insulated Gate GaAs Field Effect Transistors

Professors N.Cheung, S.Wang, C.Hu and W.Oldham with James Chan

The usage of aluminum nitride (AlN_x) as a possible dielectric material for GaAs insulated-gate field-effect transistors (IGFET), or metal-insulator-semiconductor field-effect transistors (MISFET), is being investigated. We first study the interfacial reaction of sputtered deposited AlN and III-V semiconductors. The next step is to evaluate the electrical properties of Metal/AlN/AlGaAs/GaAs device structures.

AlN_x samples have been prepared under various sputtering conditions, and numerous composition and electrical tests have been performed, indicating that AlN_x is a good insulator material.

We used an RF reactive sputterer supported by an Al target and a mixture of Ar/N_2 gases to prepare the AlN_x samples. Different sputtering conditions with varying plasma power and Ar/N_2 gas mixtures have been utilized and compared. Rutherford Backscattering composition tests were employed to analyze the composition of the dielectric films. With an Al/N ratio of 1:1 (AlN_x , with x=1.0) as the ideal goal, results show that a 115:40 Ar/N_2 mixture at a plasma power of 300 watts gives a maximum nitrogen content of x=0.8, with minimal oxygen concentration (See Figure 1).

The refractive index of the samples have also been extracted. Although index values were consistent within each run, they varied from run to run. Values ranged from 1.982 to 2.265. The average index for AlN_x samples prepared under the 115:40 Ar/N_2 gas mixture is 2.064. This compares favorably with data reported in the literature (n_f = 2.152). [1]

Current-voltage tests measured leakage currents on the order of $100pA/cm^2$ for AlN_x film sputtered on p-type Si substrate at the 115:40 Ar/N_2 gas ratio. A breakdown field on the order of 10^6 V/cm has been observed (See Figure 2), which indicate the dielectric nature of aluminum nitride.

We are in the process of finding a correlation between dielectric quality (i.e. leakage currents, breakdown field) and refractive index. In addition, test diode structures made of AlN_x film on GaAs as well as AlAs substrates are being fabricated. The AlGaAs/GaAs substrates were grown by MBE in the MBE Laboratory of UC-Berkeley (Prof. J.S. Smith). Fabrication and testing will commence early next month, at which time we will perform I-V and C-V characterizations. These tests should reveal further information concerning the interfacial quality of AlN_x on these various substrates, and its feasibility as an insulator material for GaAs MISFETs.

References

[1] Wang, D.H, and Liang, Guo, "Optical Properties of Sputtered ALN Films and Coated GaAs," *Thin Solid Films*, 158(1988), pp.39-43.

Publications

[1] Chunlin Liang, "Fabrication Technology and Device Modeling for Gallium Arsenide Metal-Semiconductor Field-Effect Transistor," Ph.D. Thesis, UC-Berkeley, 1990.

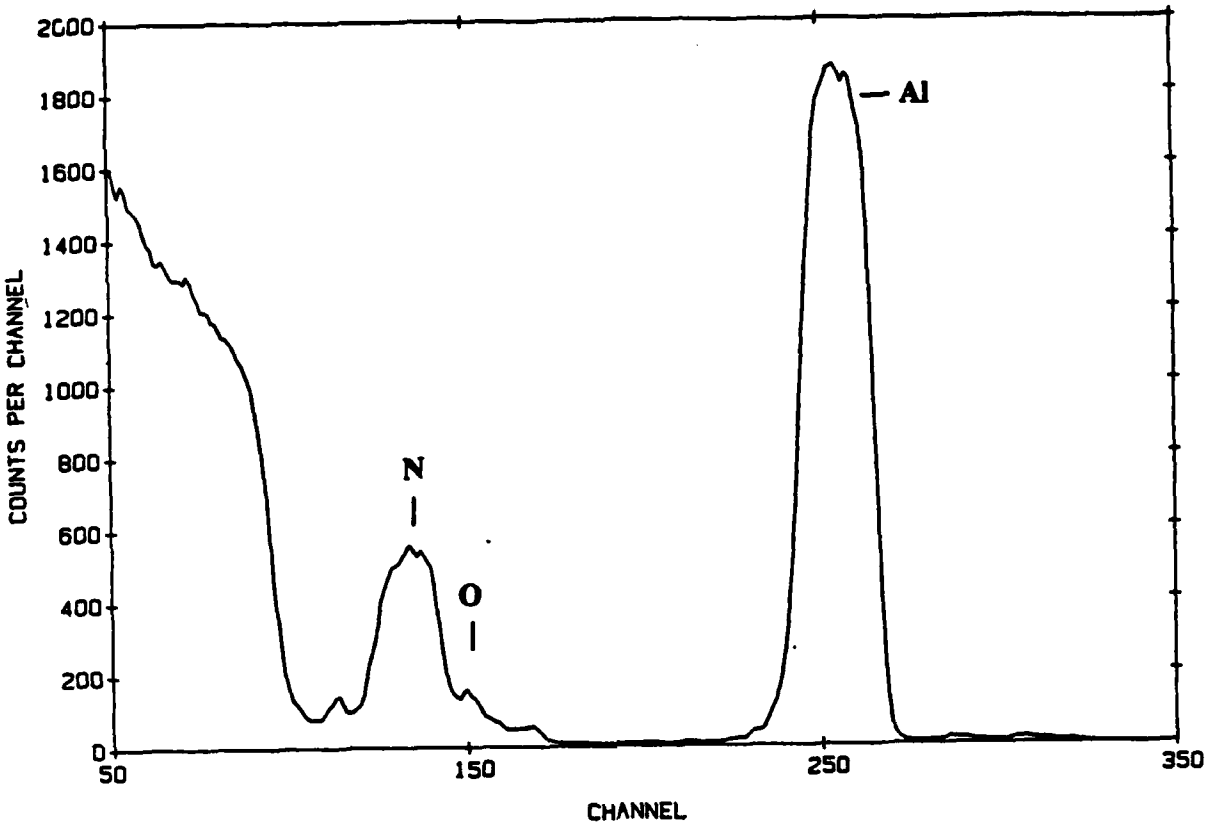


Figure 1: RBS analysis of AlN_x film sputtered at 300W, Ar/N_2 ratio of 115 sccm:40 sccm with x=0.8.

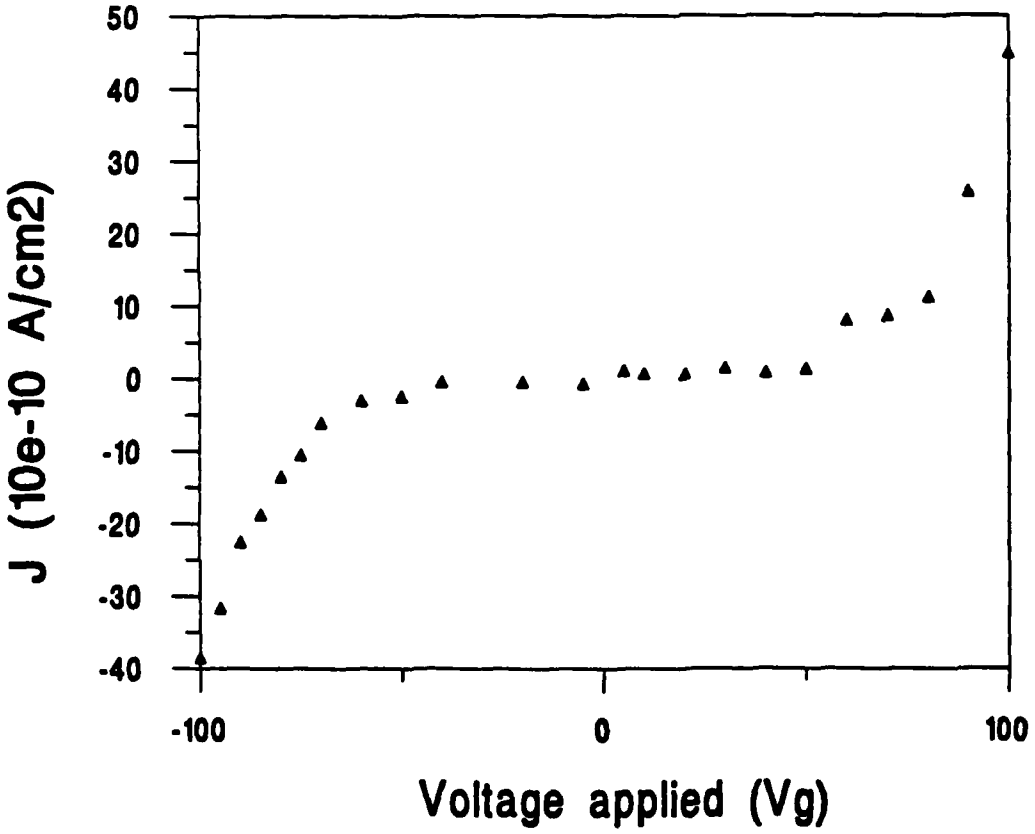


Figure 2: Current Density-Voltage characteristics for ALN_x sputtered on p-type Si with conditions described in Fig. 1. BV ~ 50V, E_{max} ~ 10^6 V/cm.

III-A. Stochastic Neural Networks and Application to Signal Processing

Professor Avideh Zakhor with Sun-Im Shih

During the last 8 months, we have continued our investigation of signal processing applications of neural networks. The particular application we are exploring is continuous phase modulation receivers. Constant envelope continuous phase modulation (CPM) schemes are important in peak power limited communication applications such as satellite transmission systems and wireless communication terminals. These schemes are generally characterized by high packing densities and prohibitively complex receiver structures. For instance, while their packing density increases with partial response L, and alphabet size M, their optimal ML receivers require a bank of matched filters whose size grows as M^L. This bank of filters is followed by a Viterbi decoder which draws heavily on computational resources. Specifically, while reducing the modulation index h improves bandwidth efficiency, the number of states in the Viterbi decoder increases with the denominator of h.

In the past few months, we have developed neural network based receiver structures for constant envelope CPM systems. Our motivation is to reduce the complexity of implementation by casting the demodulation task into the more general framework of a neural network classification task. In so doing, we replace the matched filter banks and the Viterbi decoder of the optimal receiver with a feed-forward net trained to demodulate the incoming baseband signal. Our approach is to replace the entire receiver structure, excluding timing recovery, with a multilayer feedforward neural net unit whose inputs are time samples of incoming baseband signals, and whose outputs are the decoded symbols.

We have simulated the neural net based receiver for binary 3RC with modulation index of 0.8 and found its performance to be within 3.5 dB of the optimum Viterbi based receiver at probability of error of 10^{-3}. The architectural parameters of this network are as follows: 18 input nodes, 2 samples per symbol interval and observation length of 9 symbol intervals, 30

hidden nodes and one output node. This network was trained using the well known backpropagation learning algorithm and noisy examples. We have examined the effects of architecture parameters to decoding performance of our neural net based receivers. The main conclusion is that while increasing the number of input nodes and hidden nodes improves the performance of the receiver, the training complexity grows with the size of the network. We have also developed analytical techniques to predict the signal to noise ratio performance of the network without having to simulate it. This not only enables us to understand the relationship between network parameters and performance, but also can be used as a tool in any classification application of multilayer networks.

A number of criteria can be used to compare our architecture with conventional decoders. These include implementation cost in terms of area and power, arithmetic and memory requirements, performance training time, decoding speed and training program complexity in terms of lines of code. We shall first restrict ourselves to a simplistic comparison of arithmetic complexity. A network with I input units, H hidden units and O output units requires $H\times(I+O)$ multiplies and additions per symbol interval. In a digital implementation, the Sigmoid would be evaluated by way of a table lookup, and $H+O$ lookups would be required per symbol interval. An analog implementation of the nonlinearity, is also conceivable, although this may degrade the precision of the nonlinearity. It is not clear from this what level of accuracy is required in NN architectures. Thus, our best performing NN with 36 input nodes and 30 hidden nodes requires about 1110 adds and multiplies, in addition to 31 Sigmoid evaluations per demodulated symbol.

For comparison, consider conventional CPM demodulation. The arithmetic complexity is proportional to the number of trellis states, pM^{L-1}. One can exploit the symmetry of the in-phase and quadrature components to reduce the required number of matched filters to $2M^L$, or 16 filters for binary 3RC. Suppose we use filters with 10 taps each, then 160 multiplies and

adds are required. An additional overhead of 40×2 multiplies, 40×3 adds, and 80 trigonometric evaluations to be implemented by table lookup over the 5 possible phase states are required to form the inputs to the Viterbi processor. Viterbi decoders use efficient recursive algorithms to calculate the path metrics, however they require large amounts of memory to keep track of the actual paths which will be used to ultimately make symbol decisions. At each stage, an add compare select function for each phase state is required, that is 40 adds and 40 compares. An additional, 20 compares are required to decode the symbol. An approximate total would give 320 adds, 220 multiplies, and 60 compares, in addition to 80 coarse trigonometric table look-ups per stage. When compared to the figures obtained for the NN classifier, we conclude that NN numerical complexity approximately exceeds that of a conventional implementation by a factor of 3.

In addition, a digital implementation of a NN decoder would require the storage of 1110 weights in ROM versus the 160 filter coefficients needed in a conventional receiver. However NN require virtually no RAM while a conventional receiver would need at least 220 RAM locations, to keep track of trellis paths. This is but a crude comparison of memory requirements, since the accuracy used to represent weights, filter coefficients and path metrics, has been disregarded, but it shows a tradeoff between ROM and RAM. This tradeoff is important, since RAM can take up 50 percent of the area of a custom chip for Viterbi decoding.

The NN architecture has a number of other advantages compared to the conventional receivers. The first is the regular and parallelizable nature of NN, which one might easily map to homogeneous architectures, such as systolic arrays, which are amenable to VLSI implementations. Indeed the Viterbi algorithm is an inherently serial algorithm in that previous decoded symbols are required to decode the present and future ones. Parallel implementations of the Viterbi algorithm, based on block partitioning of the data, require a synchronization period to estimate the initial state in the trellis diagram, before decoding can begin. This is in fact the

main reason for overlapping the partitioned blocks. The above-mentioned synchronization period can potentially limit the extent to which the algorithm can be parallelized. Our proposed NN implementation has no feedback, thus by using more than one network in parallel, we can make the demodulation rate arbitrarily large. A second advantage lies in the speed of demodulation. Clearly NN receivers would be faster since they require only a forward pass, while conventional decoders usually backtrack in order to obtain demodulated symbols. A third advantage, may be allowing for online adaptation to the noise characteristics of the channel, and thus providing a form of nonlinear equalization. A fourth advantage, which at this point is mere conjecture, might arise from the a favorable scaling of the NN classifier to an increase in the complexity of the modulation scheme. This is particularly important considering that conventional methods scale as M^L.

Publications/Presentations

[1] G. de Veciana and A. Zakhor, "Neural Net Based Continuous Phase Modulation Receivers," submitted for publication to *IEEE Transactions on Communications*.

[2] S. Hein and A. Zakhor, "Optimal Decoding for Data Acquisition Applications of Sigma Delta Modulators," presented at the 24th Asilomar Conference on Circuits and Sytsems, Nov. 1990.

[3] A. Zakhor, S. Hein, and K. Ibraham, "New Properties of Sigma Delta Modulators with DC Inputs," submitted for publication to *IEEE Transactions on Communications*, November, 1990.

III-B. Learning and Generalization by Neural Networks

Professor A. Sangiovanni-Vincentelli with Arlindo Oliveira and Alan Kramer

In this work unit we proposed to assess whether the use of logic synthesis techniques could be used in the specification of the interconnection patterns for neural networks architectures.

The development of a special purpose logic synthesizer, targeted for the problem of rule induction from examples has been undertaken. This logic synthesizer generates the most compact two-level network of threshold gates that matches a given input-output specification. The results have shown that a formal approach to the problem of deriving an adequate architecture for neural networks is not only feasible but highly desirable.

Apart from proving that this approach is appropriate for the derivation of appropriate rules in cases when a compact two-level representation is known to exist, this research has led to some very interesting conclusions:

1. Two level representations are not always appropriate. There are cases where either compact two-level representations do not exist or, although they do exist, they do not generate an adequate induction hypothesis. Results from hand-written character recognition have shown that although a one-gate network adequately fits the training set (for each character, in a one writter database we tested it on) the results on the test set are not perfect, with some characters being left unclassified. These results and other related ones have shown us the need for the extension of the techniques used so far to the synthesis of multi-level networks.

2. When a restriction on the largest possible size of the weights in a given network is allowed, an algorithm for the derivation of the appropriate connection weights can be derived from logic synthesis techniques. This algorithm runs in a time polynomial in the number of examples and the maximum weights size and is, therefore, much more

efficient than the ones currently used like the perceptron rule or error back-propagation. A strong convergence theorem was derived for some (somewhat restricted) conditions, although work on the relaxation of these conditions is under way. This limitation on the allowed weight sizes, far from being a burden, is usually an advantage if physical implementation of the network is desired. In fact, implementation of the large weights commonly required by alternative techniques is usually cumbersome and, in some cases, unfeasible. Our approach, on the other hand, derives networks with small integer weights, which are easier to implement.

In this work unit we also proposed to investigate massively parallel analog computation and its application to the real-world computational task of handwritten character recognition.

Close collaboration with technology groups is necessary to insure that the computing architectures we are proposing are feasible. To that end we have focused on the use of EEPROM devices as our basic computing structure. We are currently investigating several algorithms which would be efficient in implementing this technology.

The first of these is a standard feedforward neural network architecture based on novel EEPROM-based synapses. Conventional neural networks are based on synapses which perform multiplicative weighting. Once EEPROM devices are used for weight storage, the additional area needed to perform analog multiplication dominates that required to implement a synapse. We have discovered a class of highly compact, nonlinear weighting functions based on the use of the novel EEPROM I-V characteristic. This approach results in a synapse requiring one or two EEPROM devices and promises nearly a hundred-fold increase in the density of neural network implementations.

The synapses we are investigating are non-linear and thus not strictly in keeping with the mainstream in neural network learning. We have developed a neural network simulator to determine the usefulness of these non-linear synapses when applied to a real-world learning

task. We are focusing on the problem of handwritten character recognition in this work. Initial results have been promising but more investigation is needed.

The second algorithm which we are investigating is the nearest neighbor classification algorithm. This algorithm has existed for a long time and, given enough example points, has been known to perform well on vector-encoded classification tasks. It has not gained wider use because computation has been limited to a sequential digital substrate so that the distance to only one neighbor could be measured at a time. Conventional research into the use of this algorithm has focused on the development of k-d trees and other data structures to cut the nearest neighbor search down to log time, as well as ways to reduce the number of example points needed for the algorithm to perform well.

We have also discovered a novel EEPROM-based architecture for performing analog distance computation in parallel. See under the Significant Accomplishments section of this report for a discussion of this development.

Publications

[1] A. Kramer, P. K. Ko and A. Sangiovanni-Vincentelli, "Massively Parallel Analog Geometric Computation Using EEPROMs", Abstract for Neural Networks for Computing Conference, Snowbird, Utah, submitted November 1990.

[2] A. Kramer, C. Sin, R. Chu, and P. K. Ko, "Compact EEPROM-Based Weight Functions," Proceedings of the 1990 IEEE Conference, in *Neural Information Processing Systems* 3, San Mateo, California: Morgan Kauffman, March 1991.

[3] Arlindo Oliveira and Alberto Sangiovanni-Vincentelli, "Empirical Learning of Boolean Functions Using Logic Synthesis", Abstract of Neural Networks for Computing Conference, Snowbird, Utah, April, 1991. Also, Extended Abstract submitted to Conference on Learning Theory, Santa Cruz, July 1991.

III-C. Reconfigurable Analog Elements for Neural Nets

Professors Ping K. Ko and Chenming Hu

Investigation of using EEPROM devices as reconfigurable analog weights continued in the past year. Two full CMOS-EEPROM runs were completed, which demonstrated our ability to fabricate EEPROM-based CMOS circuits.

The first run finished in May produced EEPROM devices with various geometries and floating-gate structures and enabled us to verify a drain current model for EEPROM devices. The model is needed for understanding the design tradeoffs among the various substructures when a EEPROM cell is used for analog storage. Based on the modeling exercise and experimental results, we have devised a promising high-density EEPROM based analog synaspe. Other process and device information we collected from the first run also helped us get the last glitches out of the technology.

The second set of 12 wafers, which was completed in December, appears to be fully functional based on our preliminary test results. They contain several varied designs of the EEPROM-based synaspe, as well as constant-charge-packet injecting circuits based on charge-pumping or the CCD principle. We are characterizing these structures at present.

With the fabrication technology in place, we are working with several neural architecture groups building demonstrational CMOS-EEPROM-based neural network chips. Ongoing chip designs include a Layered Feedforward Neural Network chip and a Nearest Neighbor Classification chip.

Research has begun on oxide antifuse as a programmable weight device. The eventual goal is to have a small device suitable for neural chips containing millions of synapses. A mask set for test structures have been designed and the first fabrication run has been completed.

Oxide antifuses consisting of 70 *angstroms* of SiO_2 between silicon substrate and polysilicon gates were fabricated. We varied the doping concentration of the substrate and the polysilicon gates. In some structures, Al covers the antifuse oxide area to investigate the possibility of forming low-resistance filaments of Si-Al entectic. Other structures insert varying resistances between the antifuse and the probe pad capacitance to study the effect of the dumping of the capacitively stored charge and energy into the antifuse at the time of oxide breakdown.

Linear resistance that is programmable by varying the program current was observed in antifuses involving low-resistivity materials. Antifuses made with high-resistivity silicon showed exponential I-V characteristics after programming. We are in the process of characterizing these behaviors.

Publications

[1] A. Kramer, C. Sin, R. Chu, and P. Ko, "A Highly Compact Linear Weight Function Based on the Use of EEPROM," Abstract of Neural Networks for Computing Conference, Snowbird, Utah, May 1990. (Not reported under previous contract).

[2] A. Kramer, C. Sin, R. Chu, and P. Ko, "Compact EEPROM-Based Weight Functions," Proceedings of the 1990 IEEE Conference, published in *Neural Information Processing Systems* 3, San Mateo, California: Morgan Kaufman, January 1991.

III-D. Architectural Issues in Parallel Computation

Heterogeneous Architectures for ANN's

Professor C. H. Séquin with Chedsada Chinrungrueng

The overall objective of our project is to evaluate the heterogeneous architectures for artificial neural networks (ANN's). Currently we are investigating a hierarchical ANN architecture consisting of two levels: The lower level is an unsupervised competitive learning network whose task is to divide the overall task into several simpler subtasks. The upper level is a collection of networks where each one is trained for solving one of the above subtasks.

During the past six months, we have investigated competitive learning algorithms that divide a given task based on a partitioning of the input domain on which the task is defined. We have developed a new competitive learning algorithm which is a modification of the traditional adaptive k-means clustering algorithm. It divides the input domain by constructing in a continuous on-line manner a Voronoi partition around a specified number of clustering centers. This new algorithm can approximate an optimal partitioning solution with efficient adaptive learning rates, which renders it usable even in situations where the statistics of the problem task slowly vary with time.

These capabilities are achieved via two mechanisms: The first mechanism guides the partition towards an optimal solution by aiming directly at minimizing the differences between the averaged variations of each cluster. This allows the new algorithm to obtain a solution closer to the optimal value than other competitive learning algorithms in the same class. The second mechanism dynamically adjusts the learning rate based on the estimated deviation from an equilibrium state where all clusters have an equal variation of the average number of input samples that they have to cover. This allows the algorithm to learn very quickly initially or after a change in the characteristics of the problem statement. As the partition approaches an optimal solution, the learning rate decreases, which in turn allows the partition to move even closer to the optimal value. Thus the algorithm can achieve a smaller residual deviation from

an optimal value than other competitive learning algorithms.

Both of these two mechanisms are based on the necessary condition for the optimality of the k-means partition, stating that: all of the regions in the optimal k-means partition have the same *within-cluster variation* when the number of regions in the partition is large and the probability distribution generating the training samples is smooth. This within-cluster variation of any cluster is defined as the sum of the squared Euclidean distance between the pattern vectors in that cluster and the center of the cluster.

Publications

[1] C. Chinrungrueng and C. H. Séquin, "Optimal Adaptive K-Means Algorithm with Dynamic Adjustment of Learning Rate", submitted to International Joint Conference on Neural Networks, Seattle, Washington, July 1991.

Fault Tolerance in Layered ANN's
Professor C. H. Séquin with Reed Clay

Our research has been concerned with the errors resulting from defective units and faulty weights in layered feed-forward ANN's. We have explored and analyzed techniques to make these networks more robust against such failures. First, using some simple examples of pattern classification tasks and of analog function approximation, we have demonstrated that standard architectures subjected to normal backpropagation training techniques do not lead to any noteworthy fault tolerance. Additional, redundant hardware coupled with suitable new training techniques are necessary to achieve that goal. A simple and general procedure has been found that develops fault tolerance in neural networks: Failures of the type that one might expect to occur during operation are introduced at random during the training of the network, and the resulting output errors are used in a standard way for backpropagation and weight adjustment. The result of this training method is a modified internal representation that is not only more robust to the type of failures encountered in training, but which is also more tolerant of faults for which the network has not been explicitly trained.

In the context of a discrete classification task, we have demonstrated that simple training with backpropagation in the presence of various hidden unit failures can lead to fault tolerance with respect to single or multiple faults. Similar robustness can be achieved by training with multiple failures or by prolonged training with single failures. Failures in the input units are equivalent to noise on the input patterns. While training the network with hidden unit failures cannot render it robust against input failures, training with input noise can lead to some degree of fault tolerance also with respect to failures in the hidden layer.

In the context of analog function approximation tasks, we have discovered a promising approach that achieves fault tolerance by tightly controlling the fractional contribution that each hidden unit makes to the linearly summed output value. We first make the observation that the worst case output errors that can be produced by the failure of a hidden neuron can be much worse than simply the loss of the contribution of a neuron whose output goes to zero. A much larger erroneous signal can be produced when the failure drives a hidden neuron into saturation, i.e., sets its output value to one of the power supply voltages.

To counter this problem, we have investigated a new method that limits the fractional error in the output signal of a feed-forward net due to such saturated hidden unit faults in analog function approximation tasks. The number of hidden units is significantly increased, and the maximal contribution of each unit is limited to a small fraction of the net output signal. To achieve a large localized output signal, several Gaussian hidden units are moved into the same location in the input domain and the gain of the linear summing output unit is suitably adjusted. Since the contribution of each unit is equal in magnitude, there is only a modest error under any possible failure mode.

We have also explored a hierarchical approach of building ANN's to obtain more general fault tolerance to within a higher degree of accuracy for analog data approximation tasks. Several redundant subnets are used to perform the same approximation task in parallel, and a

supervisory circuit combines their outputs by eliminating the signals that fall outside some margin and by averaging the other subnets outputs.

Publications

[1] C. H. Séquin and R.D. Clay, "Fault Tolerance in Artificial Neural Networks", to appear in *Neural Networks, Concepts, Applications and Implementations*, Vol. 4, Antognetti and Milutinovic, eds., Prentice Hall, 1991. Available as *Tech Report* No. 90-031 from International Computer Science Institute, Berkeley.

[2] R. D. Clay and C. H. Séquin, "Limiting Fault-induced Output Errors in ANN's", submitted to International Joint Conference on Neural Networks, Seattle, Washington, July 1991.

Presentations

[1] C. H. Séquin, "Fault Tolerance in Feed-forward Artificial Neural Networks" International Computer Science Institute, August 21, 1990.

Parallel Computing Network and Program
Professor Abhiram Ranade with M.T. Raghunath and Robert Boothe

The goal of our research is to develop high performance, cost effective networks for interconnecting processors. We would like these networks to be suitable for running a wide variety of applications including neural network simulation and circuit simulation. The work has 2 main parts: Network Design and Understanding Parallel Program Behavior.

While network design depends upon many factors, our primary concern is packaging technology. Large networks need to be partitioned and packaged in a hierarchical manner into chips, boards, and racks. Interconnections at the lower levels of the hierarchy will be cheaper and faster. We plan to explore hybrid designs in which the lower levels of the hierarchy use a denser network than the one used at the higher levels. While it is obvious that this will improve local communication performance (e.g. within a board), we believe it will also reduce the latency for long distance communication (e.g. between racks).

We are currently developing simulators for different network architectures. We evaluate

network performance by simulating the execution of a synthetic workload which approximates the execution of real programs. We simulated multistage interconnection networks to estimate the performance improvements obtained by changing the radix, dilation, routing algorithms, etc. Results are summarized in a paper presented at the second IEEE Symposium on Parallel and Distributed Processing [1]. We are currently evaluating the performance of the recently-proposed multibutterfly network [2]. In comparison to the butterfly network, the interconnection pattern of the multibutterfly network is more complex but it is capable of achieving better latency and throughput. Our current simulations attempt to quantify this improvement in performance. Based on these results and results from simulations of other networks, we will evaluate the appropriate hybrid networks.

In order to reduce the number of detailed simulations that need to be carried out, we are defining analytical transformations that help us to extrapolate the results obtained under one set of simulation parameters to other sets of parameters.

Long memory (or communication) latency is an inevitable feature of distributed multiprocessors. Many latency tolerance and avoidance techniques have been proposed, and it is desirable to evaluate them. We are developing a simulator that will enable us to do this, and in general aid in understanding the communication behavior of parallel programs. Our simulator works at the instruction level, and can model long memory latencies that can arise in typical multiprocessor networks. The simulator is nearly completed.

Simulating computers has historically been notoriously slow. Typical parallel machine simulators at the instruction level slow down by a factor of 1,000 to 2,000. This large slow down limits the size of programs and machines that can be simulated. We have developed an innovative simulator technology that, according to preliminary measurements, allows us to reduce the slow down factor to 50-100.

This tremendous speed-up has been made possible by converting what is usually an interpretive process into direct execution. This is analogous to compilation. The program to be simulated is first compiled normally and then analyzed extensively and modified to interact with the simulator only at key points. Most instructions execute directly. Only shared memory instructions call the simulator. In a typical program more than 90% of the instructions can be executed directly, taking only a single cycle. These single cycle instruction amortize the cost of the remaining simulated instructions.

Once the simulator is completed, we will begin taking measurements of and gathering statistics on the execution behavior of large shared memory parallel programs. Such statistics are desperately needed by researchers investigating parallel computers.

References/Publications

[1] M. T. Raghunath and A. G. Ranade. "A Simulation-Based Comparison of Interconnection Networks," *Proceedings of the 2nd IEEE Symposium on Parallel and Distributed Processing*, pp. 98-103, Dallas, Texas, December 9-13, 1990.

[2] E. Upfal, "An O(log N) Deterministic Packet Routing Scheme," *Proceedings of the 21st Annual ACM Symposium on Theory of Computing*, pp. 241-250, May 1989.